# Combining Heterogeneous Databases to Detect Adverse Drug Reactions

Ying Li

Submitted in Partial Fulfillment of the Requirements for
the degree of Doctor of Philosophy
under the Executive committee of the Graduate School
of Arts and Sciences

COLUMBIA UNIVERSITY

2015

**ABSTRACT**

**Combing Heterogeneous Databases to Detect Adverse Drug Reactions**

Ying Li

Adverse drug reactions (ADRs) cause a global and substantial burden accounting for considerable mortality, morbidity and extra costs. In the United States, over 770,000 ADR related injures or deaths occur each year in hospitals, which may cost up to $5.6 million each year per hospital. Unanticipated ADRs may occur after a drug has been approved due to its use or prolonged use on large, diverse populations. Therefore, the post-marketing surveillance of drugs is essential for generating more complete drug safety profiles and for providing a decision making tool to help governmental drug administration agencies take an action on the marketed drugs. Analysis of spontaneous reports of suspected ADRs has traditionally served as a valuable tool in pharmacovigilance. However, because of well-known limitations of spontaneous reports, observational healthcare data, such as electronic health records (EHRs) and administrative claims data, are starting to be used to complement the spontaneous reporting system. Synthesizing ADR evidence from multiple data sources has been conducted by human experts on an at hoc basis. However, the amount of data from both spontaneous reporting systems (SRSs) and observational healthcare databases is growing exponentially. The revolution in the ability of machines to access, process, and mine databases, making it advantageous to develop an automatic system to obtain integrated evidence by combining them.

Towards this goal, this dissertation proposes a framework consisting of three components that generates signal scores based on data an EHR system and of an SRS system, and then integrates two signal scores into a composite one. The first component is a data-driven and regression-based method that aims to alleviate confounding effect and detect ADR based on EHRs. The results demonstrate that this component achieves comparable or slightly higher accuracy than those trained with experts and existing automatic methods. The second component is also a data-driven and regression-based method that aims to reduce the effect of confounding by co-medication and confounding by indication using primary suspected, secondary suspected, concomitant medications and indications on the basis of a SRS. This study demonstrates that it could accomplish comparable or slightly better accuracy than the cutting edge algorithm Gamma Poisson Shrinkage (GPS), which uses primary suspected medications only. The third component is a computational integration method that normalizes signal scores from each data source and integrates them into a composite signal score. The results achieved by the method demonstrate that the combined ADR evidence achieve better accuracy of drug-ADR detection than individual systems based on either an SRS or an EHR. Furthermore, component three is explored as a tool to assist clinical assessors in pharmacovigilance practice.

The research presented in this dissertation has produced several novel insights and provided new solutions towards the challenging problem of pharmacovigilance. The method of reducing confounding effect can be generalizable to other EHR systems and the method for integrating ADR evidence can be generalizable to include other data sources. In conclusion, this dissertation develops a method to reduce confounding effect in both EHRs and SRSs, and a combined system to synthesize evidence, which could potentially unveil drug safety profiles and novel adverse events in a timely fashion.

## TABLE OF CONTENTS

**LIST OF FIGURES**

**LIST OF TABLES**

**LIST OF EQUATIONS**

**LIST OF ABBREVIATIONS**

ADR - Adverse Drug Reaction

ALI - Acute Liver Injury

ALS - Abnormal Laboratory Signals

AMI - Acute Myocardial Infarction

ARF - Acute Renal Failure

ASTER - Adverse Drug Events Spontaneous Triggered Event Reporting

AUC - Area under a Receiver Operating Characteristic Curve

BBC - British Broadcasting Corporation

BCPNN - Bayesian Confidence Propagation Neural Network

BLR - Bayesian Logistic Regression

CUI - Concept Unique Identifiers

CUMC/NYPH - Columbia University Medical Center/New York Presbyterian Hospital

DPA - Disproportionality Analyses

EBGM - Geometric Mean of Empirical Bayes Posterior Distribution

EHR - Electronic Healthcare Record

EMEA - European Agency for the Evaluation of Medical Products

FAERS - FDA adverse event reporting system

FDA - Food and Drug Administration

FDR - False Discovery Rate

GIB - Upper Gastrointestinal Bleeding

GPS - Gamma Poisson Shrinker

IC - Information Component

ICD-9 - International Classification of Disease, Version 9

ICTPD - Information Component Temporal Pattern Discovery

LASSO - Least Absolute Shrinkage and Selection Operator

LLR - Lasso Logistic Regression

MGPS - Multi-item Gamma Poisson Shrinker

MI - Myocardial Infarction

MQIC - Medical Quality Improvement Consortium

MeSH - Medical Subject Headings

MedDRA - Medical Dictionary for Regulatory Activities

MedLEE - Medical Language Extraction and Encoding

NLM - National Library of Medicine

NLP - Natural Language Processing

OMOP - Observational Medical Outcomes Partnership

OR - Odds Ratio

PEM - Prescription Event Monitoring

PS - Propensity Score

PSM - Propensity Score Method

PhV - Pharmacovigilance

RCT - Randomized Controlled Trial

RF - Risk Factors

ROC - Receiver Operating Characteristic

ROR - Reporting Odds Ratio

RRR - Relative Reporting Ratio

SCC - Self Control Cohort

SCCS - Self Controlled Case Series

SIDER - Side Effect Resource

SRS - Spontaneous Reporting System

STEMI - ST-elevation myocardial infarction

STITCH - Search Tool for Interactions of Chemicals

UMLS - Unified Medical Language System

VAERS - Vaccine Adverse Event Reporting System

WHO - World Health Organization

**ACKNOWLEDGEMENTS**

There are many people I must thank for their guidance and support during this dissertation. First, I must acknowledge my primary advisor, Dr. Carol Friedman, for her invaluable guidance and mentorship. Her dedication, patience and persistence have led me to become a thorough scientist.

I would also like to thank all my committee members. Dr. Ying Wei, who served as a co-advisor and actively participated in all of work presented in this dissertation. I want to thank her for her critical advice on developing statistical methods and encouragement even during difficult times. Dr. Herb Chase, who tirelessly educated me clinical knowledge on drug's side effect. Dr. Hua Xu, who shared his research experience with me and made available the data from Vanderbilt University. Dr. George Hripcsak, whose scientific and creative spirit always inspired me.

I wish to thank Dr. Patrick Ryan for brainstorming discussions that helped me rethink the research in broader and more significant ways, and for his generous help in accessing data from Observational Medical Outcomes Partnership. Dr. Santiago Vilar, for creating reference standards for evaluating proposed methods. Dr. Hojjat Salmasian, for useful discussions. Lyudmila Shagina, for her technical assistance.

Thanks also to DBMI for all the support it has provided me through the years. A special thanks to my colleagues and friends who have made these years enjoyable – Lorena Carlo, Nicole Weiskopf, Adler Perotte, Shaodian Zhang, Sharon Gorman, Rave Harpaz and Lixia Yao.

I must express my gratitude to Lei, my husband, for his continued support and encouragement. I must thank my family for all their endless love and support, in particular my father, who always comforted me when I experienced all of the ups and downs of my research.

This research has been supported by grant LM010016 awarded to Dr. Carol Friedman.

# CHAPTER 1 Introduction

## 1.1 Problem and Significance

It is perhaps a fundamental truth in medicine that there is no medication that is without risk (Coloma 2012). Even with the most rigorous efforts in drug approval and regulation, unanticipated adverse drug reactions (ADRs) may occur. In the 1960s, the thalidomide disaster affected nearly 10,000 people around the globe. Post-marketing surveillance, also referred to as pharmacovigilance, has drawn a great deal of attention from the public ever since.

The burden of ADRs worldwide is high, accounting for considerable morbidity, mortality, and extra costs. The Institute of Medicine reported in January of 2000 that an estimated 7,000 deaths per year occur due to ADRs (Kohn, Corrigan et al. 2000). Another study conducted based on hospitalized patient populations estimate that 6.7% of hospitalized patients had a serious adverse drug reaction with a mortality rate of 0.32% (Lazarou, Pomeranz et al. 1998). Also, it was estimated that over 350,000 ADRs occur in U.S. nursing homes each year (Gurwitz, Field et al. 2000). Meanwhile, one estimate of the cost of drug-related morbidity and mortality is $136 billion annually, which is more than the total cost of cardiovascular or diabetic care in the United States (Johnson and Bootman 1995). In addition, studies indicated that national hospital expenses to treat patients who suffer ADRs are estimated between $1.56 and $5.6 billion annually (Bates, Cullen et al. 1995, Bates, Spell et al. 1997, Raschke, Gollihare et al. 1998, Thomas, Studdert et al. 1999).

## 1.2 Challenges

Randomized controlled trials (RCTs) are considered to be the "gold standard" for determining a cause-and-effect relationship between a medication and an outcome, however, these trials are rarely large enough to accurately measure infrequent adverse outcomes (Black 1996). Once drugs are on the market, they are used on a much larger and more diverse population, often with prolonged periods and sometimes with a wider range of therapeutic indications (Amery 1999, Berlin, Glasser et al. 2008), and consequently unanticipated ADRs may occur. Therefore, post-marketing surveillance of approved drugs is essential for generating more complete drug safety profiles. Extending resources to observational data and methods represent a set of complementary approaches that could potentially augment ADR detection (Olsson 1998, Ahmad 2003).

Post-marketing drug safety surveillance has traditionally been conducted by systematic manual review of reports of suspected ADRs in spontaneous reporting systems (SRSs), which are mainly described by healthcare professionals, consumers, and pharmaceutical manufacturers (Hauben, Madigan et al. 2005). It is impractical to manually review the reports due to a large amount of reports as well as the continuous influx of new drugs. For the past decades, various automatic signal detection methods have been developed on the basis of SRSs to supplement qualitative clinical evaluation (van Puijenbroek, Bate et al. 2002). The success of current pharmacovigilance systems, however, is hampered by limitations inherent in the SRS databases, such as underreporting and the pitfalls of automatic signal detection methods, such as not appropriately dealing with confounding. It has shown that ADRs may be detected and acted upon too late (Topol 2004). The withdrawal of rofecoxib, together with other significant safety issues, when millions of persons have already been exposed, have stimulated initiatives worldwide to explore

new methods to facilitate earlier detection of novel ADRs. A recent resource involves mining of observational healthcare data, including routinely-collected, longitudinal electronic healthcare records (EHRs) and longitudinal billing oriented claims data. Different research groups have carried out considerable studies on the basis of large-scale EHRs or claims data and have demonstrated that observational healthcare database can augment existing pharmacovigilance systems (Coloma 2012).

Given the relative maturity of the pharmacovigilance based on SRS, the rapid development of ADR detection on the basis of observational healthcare data, and vast improvements in computing capabilities, the time is ripe to develop methods for integrating ADR evidence from two or more resources. Towards this goal, we develop a method to synergistically combine ADR signals mined from complementary data sources and demonstrate the potential of the method using a published reference standards. The quality of signals generated by the combination system depends on the quality of their counterparts produced by each individual source. It is well known that confounding effect is one of the most challenging problems leading to high false positive rates, therefore we developed two methods for controlling complex confounding effect in the EHR and SRS,

### 1.3 Research Hypotheses

This dissertation has three aims. First, to develop novel methods for alleviating confounding effect in observational health data, such as EHR, so that EHR-based pharmacovigilance method can be improved. Second, to develop novel methods for reducing confounding by indication and concomitant medications based on SRSs. Third, to develop novel methods that leverage ADR

evidence from multiple databases synergistically so that the combined method could detect ADRs more effectively than the individual data sources.

Specifically:

1. Detecting ADR signals from the observational healthcare databases, such as a single facility EHR, is challenging due to the existence of complex confounding effect that leads to the high false positive rate. The proposed data-driven and regression-based method could effectively reduce the confounding issue resulting in the improvement of ADR detection accuracy compared with other existing methods.

2. Leveraging information of concomitant medications and indications in FDA adverse event reporting system (FAERS) can improve ADR detection performance compared with the traditional measurements produced by disproportionality analyses, which are solely based on information of primary suspected medications.

3. Computationally integrating ADR evidence generated by the FDA adverse event reporting system (FAERS) and observational healthcare data can result in a more accurate and sensitive ADR detection system than systems based on individual sources.

### 1.4 Thesis Overview

In this dissertation, we develop an integrative system synthesizing ADR evidence from multiple heterogeneous databases, which includes the following components: (1) a data-driven and regression-based method for reducing confounding effect and therefore improving ADR signal detection in NYP/CUMC EHR; (2) a data-driven and regression-based method leveraging primary suspected medications, concomitant medications and indications, and alleviating confounding effect in FAERS. (3) a computational method to automatically integrate evidence

on the basis of data from observational healthcare data and FAERS, which can serve as a tool for clinical assessors in actual pharmacovigilance practice.

Chapter 2 contains background material associated with pharmacovigilance, including a) a survey of current databases, algorithms and reference standards used for post-marketing surveillance of drug safety, b) a review of relevant studies, and c) a summary of related techniques including natural language processing and biomedical terminologies.

Chapter 3 describes a data driven method to detect ADR signals using primarily inpatient data associated with a single hospital visit as well as evaluation of the method based on a reference standard consisting of two serious ADRs and drugs known to cause them. The method includes the following five steps: 1) data collection and preparation; 2) identification of candidate drug safety signals; 3) identification of confounders for specific medications; 4) estimation of the medication–ADR associations adjusting for potential confounders; 5) determination of the adjusted medication-ADR signals. The evaluation involves a reference standard consisted of 1,055 known positive drugs for two serious ADRs, and focuses on the precision of detecting known drug-ADR signals and on comparison with other existing methods using the precision as an assessment metric.

Chapter 4 presents a study of the effect of data characteristics on ADR detection methods when the resource is FAERS. In this work, we explore the use of concomitant medication and indication information in addition to primary suspected information to improve the performance of ADR detection. For evaluation, a reference standard comprising 165 positive and 234 negative drug-ADR pairs is utilized and the major assessment metric is the area under a receiver operating characteristic curve (AUC).

Chapter 5 develops a computational method to combine signals from observational healthcare databases and FAERS. In this work, we conduct three experiments involving combining FAERS with a single facility small-scale EHR, a larger-scale network-based EHR, and a much larger-scale healthcare claims database. The evaluation uses a reference standard comprising 165 positive and 234 negative drug-ADR pairs, and focuses on the AUC. Furthermore, we demonstrate that the proposed system can serve as a tool for synthesizing ADR evidence under two different scenarios that generally occur in actual pharmacovigilance practice, namely when two data sources provide either consistent or inconsistent information about particular drug-ADR pairs.

Chapter 6 summarizes and discusses the contributions and significance of the overall framework for reducing confounding effect, generating ADR signals and integrating ADR evidence, and presents the limitations, future work and overall conclusions.

# CHAPTER 2 Background

In this chapter, we provide an overview of the key data sources for pharmacovigilance, survey methods that are state of the art, describe reference standards centering on these data. We also review relevant work concerning synthesizing evidence from multiple data sources, and describe related techniques for conducting studies discussed in this thesis.

## 2.1 Overview

Pharmacovigilance (PhV), also referred to as drug safety surveillance, has been defined as "the pharmacological science relating to the detection, assessment, understanding and prevention of adverse effects, particularly long term and short term side effects of medicines". The collection of PhV information starts at the pre-approval stage, such as phase I-III of clinical trials, and continues in the post-approval stage and throughout a drug's life on the market. Typical databases used in the post-marketing stages include spontaneous reporting systems, observational healthcare databases and prescription event monitoring databases. More recently, biomedical literature and data produced by the Internet have caught researchers' attention. With a rapid increase of the data size, automatic methods to deal with data and generate ADR signals have been studied and developed.

## 2.2 Data source used in support of Pharmacovigilance

### 2.2.1 Spontaneous reporting systems

In the aftermath of the thalidomide tragedy in the late1960s, the United States Food and Drug Administration (FDA), the World Health Organization (WHO), and the European Agency for the

Evaluation of Medical Products (EMEA) and other governmental drug administration agencies independently set up spontaneous reporting systems (SRSs) designated for the collection and subsequent analysis of post-marketing safety information (Coloma, Trifirò et al. 2013). In the United States, the FDA Adverse Event Reporting System (FAERS) is the primary surveillance database used for the identification of safety problems of marketed drugs. Since its inception on 1969, a list of drugs has been removed from the market or restricted to special requirements for prescription due to safety problems, representing 1% of marketed drugs (Wysowski and Swartz 2005). Currently, FAERS contains over 5 million spontaneous reports of suspected ADRs, and receives an average of 300,000 reports per year among which the majority - 66% - come from the US (Coloma, Trifirò et al. 2013).

Most of case reports collected by the SRS centers are either required to be submitted by pharmaceutical companies, or are voluntarily reported by healthcare professionals and consumers. Each report usually includes one or more adverse events that appear to be associated with the administration of a drug; in addition, concomitant medications, indications and limited demographic information are also reported. Although case reports submitted to the SRSs do not necessarily imply causal relationships, the scenario of multiple reports which are similar and which independently originate from different sources raises the degree of suspicion, and sometimes have been considered sufficient for regulatory decisions (Brewer and Colditz 1999). SRSs can be effective in revealing unusual or rare adverse events that occur with the initial use or short-term use of medications. For example, methods using an SRS rapidly identified that temafloxacin was the cause for the ADR hemolytic anemia in otherwise healthy individuals because hemolytic anemia was rare in the general population and occurred within 1 week of drug use (Blum, Graham et al. 1994). However, SRSs do not rapidly lead to ADR detection if the

adverse event is relatively common but not necessarily drug-related in the general population such as approximately 30-year gap between the detection of autoimmune like disorders attributable to breast implant and its initial use(Kessler, Natanblut et al. 1993, Sanchez-Guerrero, Colditz et al. 1995). The autoimmune like symptoms are relatively common in women without implants and the recognition of this ADR is subjective, leading to underreporting when physicians and patients lack the knowledge of connecting breast implant to autoimmune like symptoms. Additional limitations of SRS include biased reporting influenced by media coverage or the length of time on market (Eberth, Kline et al. 2014), incomplete, inaccurate and duplicate reporting. For example, a study showed that patients were less likely to attribute an ADR to the prescribed medication than an expert panel that reviewed the event forms (Mitchell, Henry et al. 1988), sampling biases whereas all the reports are related to corresponding ADRs so that information on the number of patients who take a drug of interest but do not develop an ADR is unknown (Brewer and Colditz 1999, Bate and Evans 2009), and duplicate reporting whereas multiple reports referring to the same adverse events are collected from different sources such as consumers, drug manufacturers and investigators (Sakaeda, Tamon et al. 2013).

### 2.2.2 Observational healthcare data

Based on the forgoing discussion, it is apparent that one of greatest limitations in the SRSs to post-marketing safety surveillance is their passive property and therefore delaying ADR detection. The imperative to shift the paradigm toward a more proactive approach calls for the attention of regulators and researchers (Gagne, Glynn et al. 2012). A proactive approach is a procedure that actively and routinely screen the data collected during the routine clinical care in order to generate hypothesis about the association between certain medications and selected

ADRs. With the advance in information technology and increasing adoption across the world, electronic health records incorporated with detailed clinical data has become potential resources for proactive ADR detection (Psaty and Burke 2006, Stratton, Baciu et al. 2007, Platt, Wilson et al. 2009).

An EHR is a longitudinal electronic record of patient information generated by one or more encounters in routine clinical care. This record usually includes structured information such as laboratory test results, medication orders and diagnostic codes for billing, and unstructured information in narrative text such as patients' signs and symptoms, disease status and severity, and medical history. The EHR is initially designed and implemented to trace accurate, up-to-date, and complete information about patients at the point of care. Nowadays, clinical researchers are increasingly interested in the secondary use of clinical data, which are promising for comparative effectiveness research, outcomes research, epidemiology, public health research and drug surveillance (Hersh 2007, Safran, Bloomrosen et al. 2007). The function of drug surveillance based on EHR was exemplified by an initial pilot project by Partners Healthcare - Adverse Drug Events Spontaneous Triggered Event Reporting (ASTER), which allowed selected physicians to report suspected ADRs in an automated way. The system prompted an alert for reporting when the physician indicated in the EHR that a drug had been discontinued because of an adverse event (Linder, Haas et al. 2010). However, upon the evaluation, they found that most of the ADRs captured and reported to FDA are known events, for example, ADRs that are included in product labeling, for the suspect drugs (Brajovic, Piazza-Hepp et al. 2012). In terms of discovering novel ADRs, a study used the UK primary care databases IMS disease Analyzer MediPlus to show how longitudinal data may facilitate early signal detection (Bate, Edwards et al. 2004). Several studies showed the earlier detection of cardiovascular events associated with

the use of cyclooxygenase-2 inhibitors (coxibs) within an EHR database (Curtis, Cheng et al. 2008). Moreover, a research explicitly demonstrated that when data were restricted to time prior to a regulatory action, the potential signals were much stronger when using the EHR than using SRSs (Patadia, Schuemie et al. 2014). However, EHRs introduce other challenges. First, usually only researchers affiliated with a medical center can access clinical notes within the institution even when they are de-identified (Friedman, Rindflesch et al. 2013), and consequently jeopardize the procedure for accumulating data from multi-site medical centers in order to detect rare events or study newly-marketed medications. Second, the data collection procedure and data quality of EHR vary across different providers and hospitals. Third, medical records usually mention the patient's medications, symptoms, diseases, and procedures individually without mentioning their relationships. Fourth, most of the information is buried in narrative clinical notes, and is inaccessible for automated applications.

Similar to EHRs, linked administrative databases, such as Medicare and commercial healthcare claims databases, are emerging as a source for ADR detection. In comparison to EHRs, structured claims data, consisting of diagnosis codes, procedure codes and prescriptions, have relatively low sensitivity for detecting ADRs, weaker coverage of symptomatology, and are vulnerable to inaccuracies as they are oriented toward billing(Nadkarni 2010, Classen, Resar et al. 2011).

### 2.2.3 Prescription event monitoring databases

Prescription-event monitoring (PEM) was first suggested 25 years ago as a way to monitor the overall safety of newly marketed medicines as used in real-life clinical practice, usually in cohorts of at least 10,000 patients. The number of 10,000 patients was chosen since it is

estimated that a sample size of 10,000 patients should detect at least three ADRs with 85% power even when ADR occurs at a rate of 1 in 2000 and assuming the background rate is zero (Strom 2011). The United Kingdom was the first country to adopt PEM which actively solicit information of suspected ADRs involving demography, indication, dose, reason for stopping medication (if applicable), any events that had occurred since starting medication, whether any events were suspected to be ADRs and whether events were reported to the UK Regulatory Authority or manufacturer (Rawson, Pearce et al. 1990) (Bate and Evans 2009). A similar system called Intensive Medicines Monitoring Programme is carried out in New Zealand which monitors the first 10,000 patients exposed to a new drug for a mean of almost five years (Coulter 2000).

In general, prescription follow-up information provides a denominator - the number of patient exposed and a numerator - the number of ADRs - for calculating ADR rates. Reporting rates are hence much higher than voluntary reporting. An example of ADRs identified by PEM include cough with captopril (Coulter and Edwards 1987). However, since PEM only contains details of clusters of patients exposed to a particular drug, the lack of an adequate control group is a limitation. For example, tolterodine did not show evidence of hallucinations as an ADR because the control group contained patients prescribed other drugs known to cause hallucinations. When the data from these patients were removed, an ADE signal for tolterodine was discovered (Heeley, Wilton et al. 2002).

### 2.2.4 Other promising data sources for pharmacovigilance

Systematic review of biomedical literature is a comprehensive scientific evidence source to confirm or reject a possible drug-ADR causal relationship. Shetty et al expedited this process by

retrieving possible ADR case reports from MEDLINE on the basis of National Library of Medicine's (NLM) Medical Subject Headings (MeSH) index and a Lasso-based document relevance classifier, and then applied a disproportionality analysis to identify statistically significant drug-ADR associations(Shetty and Dalal 2011). Avillach et al. devised an ADR identification process based entirely on MeSH annotations. The MeSH subheadings of 'chemically induced', 'adverse effects' and 'pharmacological action' were used to link drugs and medical conditions in an article as candidate drug-ADR pairs. They then identified a possible drug-ADR association by using a threshold of three articles whose MeSH annotations contained the studied drug-ADR pair (Avillach, Dufour et al. 2013). In contrast, Wang et al developed a machine learning approach based on the text of the article from PubMed to support pharmacovigilance for particular ADRs they were interested in (Wang, Haerian et al. 2011).

User-posted data on social media has become a useful resource for ADR monitoring. In terms of sources, both health-related sites, such as PatientsLikeMe and DailyStrength, and general social media data, such as Twitter, have been used for ADR detection. In a recent paper, Freifeld et al. described an analysis of Twitter posts for references to drugs and adverse events, with comparison to reporting patterns in the US FDA FAERS and showed that the Spearman rank correlation rho of 0.75 ($p < 0.0001$) between Proto-AEs reported in Twitter and FAERS by SOC (Freifeld, Brownstein et al. 2014). Health-related sources tend to contain higher proportions of relevant data while the amount of data from general social media websites is significantly larger.

In terms of methods for detecting ADRs, Medawar et al. initiated a study in 2001, which validated a relationship between suicidal thoughts and the antidepressant paroxetine by reviewing posts to an online discussion board and emails sent to a major British Broadcasting Corporation (BBC)-TV documentary programme (Medawar, Herxheimer et al. 2002). Lately,

supervised classification techniques for detecting posts associated with ADR mentions, and lexicon-based approaches for extracting ADR mentions from texts have become popular (Sarker, Ginn et al. 2015). In parallel, the Internet search patterns have been explored for similar purposes. For example, White et al. conducted two studies to examine the feasibility of a signal detection system based on the web search logs (White, Tatonetti et al. 2013, White, Harpaz et al. 2014).

### 2.3 Methods applied in pharmacovigilance

#### 2.3.1 Disproportionality analysis

Disproportionality analyses (DPA) are routinely applied to SRSs (A. Bate et al., 1998; W. DuMouchel & Pregibon, 2001; W DuMouchel, 1999; Lindquist et al., 1999; Lindquist, Stahl, Bate, Edwards, & Meyboom, 2000; Noren, Bate, Orre, & Edwards, 2006) to measure the strength of reported drug-event associations. DPA involves calculating surrogate observed-to-expected ratios in which each drug-ADR pair is compared to background across all other drugs and events in the database. Two of the most widely cited measurements are the relative reporting ratio (RRR) and reporting odds ratio (ROR). RRR is the ratio between the number of reports concerning a particular drug-ADR combination to an expected number under the assumption that the drug and ADR occur independently (Norén, Hopstadius et al. 2013). ROR considers SRS as source data for a case-control study, under the assumption that the odds of the ADR are not affected by the drug (Rothman, Lanes et al. 2004). Both RRR and ROR do not address the sampling variance issue. Multi-item Gamma Poisson Shrinker (MGPS) and Bayesian confidence propagation neural network (BCPNN, information component (IC) is the statistical score) adopt Bayesian approaches to cope with sampling variance by shrinking RRR or IC towards a prior

when less data concerning the drug-ADR pair is available (DuMouchel 1999). MGPS method is the routine ADR detection algorithm used in the FDA FAERS, and BCPNN used to be the routine ADR detection method applied in the WHO VigiBase, which was replaced by a much simpler method developed by Noren et al recently (Norén, Hopstadius et al. 2013). Lately, the DPA method was adapted to take temporal information into account to measure the drug-ADR associations in observational healthcare databases, such as healthcare claims database and structured electronic health records (Schuemie 2011, Zorych, Madigan et al. 2013). Meanwhile, the DPA method was applied to measure the drug-ADR association on the basis of ADR case reports in the MEDLINE database (Shetty and Dalal 2011).

However, all the above methods measure lower order associations, such as a single drug-ADR pair without considering the effect of confounding factors. A confounder is an extraneous variable, either observed or unobserved, that mediates an association between two other variables. For example, alcoholism is a confounder that could lead to a suspicious relationship between the medication Naltrexone and pancreatitis because Naltrexone treats alcoholism, which often leads to pancreatitis. If not properly accounted for, confounding may lead to the discovery of suspicious associations and therefore erroneous study conclusions.

### 2.3.2 Multiple regressions

Randomization is an experimental design to randomly allocate subjects to the treatment group and other control groups so that the groups have similar distributions of age, gender, behaviors, and virtually all known and unknown possible confounding factors. The data collected by the randomization design are supposed to be free of confounding effect. However, as in the case of SRS and observational healthcare databases where data have already been collected, the

characteristics of patients in exposure or unexposed group could not be balanced through randomization, confounding should be addressed in the analysis stage. Stratification is a standard procedure to alleviate confounding effects but it is not effective in situations where a large amount of potential confounders need to be examined. A more appropriate approach to handling confounding is by the use of multiple logistic regression, or new extensions of logistic regression to very-large-dimensional data, known as regularized or Bayesian logistic regression (BLR). Caster et al. described an application of BLR to the WHO SRS, involving an attempt to address confounding caused by co-medication and a "masking" effect (Caster, Norén et al. 2010). Masking effect is the suppression of a statistical reporting association between a drug and an adverse event due to large numbers of reports for that adverse event in connection with another drug or drugs (Wang, Hochberg et al. 2010). For example, the association between the anti-depressive drug venlafaxine and the ADR rhabdomyolysis were masked by media focus on the withdrawal of a drug (cerivastatin) causing rhabdomyolysis (Caster, Norén et al. 2010). Later on, regularized logistic regression is applied to the healthcare claims databases and EHRs to eliminate confounders (Harpaz, Haerian et al. 2010, Ryan, Madigan et al. 2012, Li, Salmasian et al. 2013). Propensity score (PS) method is another commonly used regression-based analytic approach for controlling confounding in the analysis stage (Rosenbaum and Rubin 1983, Patrick, Schneeweiss et al. 2011). Propensity scores combine information from a large number of covariates into a single variable representing a subject's probability of receiving a particular treatment, given the measured characteristics. This score can be used for matching, stratification, as a weighting factor, or as an adjustment factor in multivariable regression (Stürmer, Joshi et al. 2006). Tatonetti et al used PS method to identify potential drug-drug interaction between paroxetine and pravastatin that could possibly cause hyperglycemia (Tatonetti, Ye et al. 2012).

### 2.3.3 Epidemiology design

Epidemiological methods, including cohort, case-control and self-controlled designs, have been frequently applied to observational healthcare data. Cohort design identifies two subgroups of the population on the basis of the presence or absence of the exposure (Rothman, Greenland et al. 2008). The non-exposure group could consist of patients who did not take particular medications or who took other medications whose indication is the same to the studied medication. The association is measured by comparing the presence and the absence of the outcome between two groups. A case-control study has the same specifications as a cohort study, except that the roles of exposure and diseases are reversed (Rothman, Greenland et al. 2008). The case group consists of patients developing the disease and the control groups consist of patients who are free of the disease. The relationship is measured by the presence and absence of exposure for individuals in both groups. The main advantage of case–control studies as compared with alternative study designs such as cohort designs is their data efficiency, which permits the study of rare events (Jewell 2003). Self-controlled design can produce results that are statistically and clinically valid with far fewer patients than would otherwise be required by using each patient as his or her own control. The self-controlled case series (SCCS) is a type of self- controlled design which assumes that ADRs arise according to a nonhomogeneous Poisson process, with each subject having an individual baseline of non-exposure event rate that is constant over time, and with periods of exposure resulting in a multiplicative effect on the baseline rate (Simpson 2011). The above three designs were intensively examined by OMOP on the basis of five databases. In an experiment conducted by OMOP, high dimensional propensity score based cohort study achieved a sensitivity of 56%, specificity of 82%, and positive predictive value of 38% in the detection of 53 associations corresponding to true ADEs and

negative controls. The implementation of a case–control design achieved close to 100% sensitivity, but at the expense of extremely low specificity of 15% (Ryan, Madigan et al. 2012). Self- controlled methods, such as self-controlled case series, temporal pattern discovery and self-controlled cohort, had higher predictive accuracy than cohort and case–control methods across all databases and outcomes. However, the distributions of point estimates across all analysis methods for the negative controls, which are supposed to be centered on zero, were positively biased. (Ryan, Stang et al. 2013).

### 2.3.4 Unsupervised machine-learning methods

Unsupervised machine-learning approaches, such as clustering, association rule mining and network analysis, have been used for the identification of more complex or higher-dimensional drug safety phenomena as well as for data abstraction and pattern discovery. In general, the clustering algorithms could be used to group patients with similar symptoms or diagnoses, which segment a large patient population to a smaller set of specific homogeneous subgroups (clusters) without losing much information about the whole population. The drug-ADR associations could further be calculated within these relatively homogeneous clusters and summarized using techniques such as Cochran-Mantel-Haenszel method. Because of the heterogeneity between clusters, this analysis can also be helpful in hypothesis development about the nature of the variation between subgroups. For example, if a database contained details of different cardiac pathologies (e.g. valvular heart disease) and medication (e.g. fenfluramine-phentermine), clustering analysis may have segregated patients according to heart disease and identified fenfluramine-phentermine as one of the main factors in this group. We could then explore the

hypothesis of an association or causal link between cardiac valvular disease and fenfluramine-phentermine (Wilson, Thabane et al. 2004).

### 2.4 Integration of ADR evidence from heterogeneous databases

Regulatory decision-making based on integrating available research data from multiple data sources to determine whether a drug is safe is a complex process (Anello and O'neill 1996). The main use case for leveraging multiple data modalities is to improve signal detection via evidence combination. In this regard, the questions that need to be studied are whether we should use some data sources for hypothesis generation while reserving others for confirmation, or combine data sources in a novel way to generate hypothesis.

Tatonetti *et al*. discovered a potentially new drug-drug interaction, which can lead to unexpected increases in blood glucose levels, between paroxetine and pravastatin based on SRS, and then validated this interaction using multi-center EHRs (Tatonetti, Denny et al. 2011, Tatonetti, Fernald et al. 2012). Duke et al. predicted probable novel myopathy-associated drug-drug interactions based on the literature, and evaluated them using a large EHR database(Duke, Han et al. 2012). Xu et al boosted drug-ADR pairs' signals generated from FAERS by incorporating the information about their MEDLINE occurrences. The key assumption in their study was that if a drug-ADR pair appears in both MEDLINE and FAERS database, then this pair likely has a true ADR relationship and if this pair also appears in FAERS many times, then the probability of it being a true "drug CAUSE ADR" pair is high (Xu and Wang 2014). Harpaz et al. claimed that a combinatorial investigation of SRS and the EHRs either lead to increased evidence or statistical power of findings, or would facilitate new discoveries that may not be possible with either source separately (Harpaz, Vilar et al. 2012). In particular, the study analyzed 4 million reports obtained

from FAERS together with information extracted from 1.2 million EHR narratives using disproportionality analysis to generate a list of ADRs and then re-ranked them on the basis of signal strength calculated from the EHR. The results showed that the accuracy of signal detection, measured by the 'Precision at K' metric (Baeza-Yates and Ribeiro-Neto 1999, Liu and Zsu 2009), was improved. A reference standard of three serious adverse reactions and over 600 established and plausible ADRs was used to evaluate the proposed approach against the single FAERS-based signal detection system. Established ADRs are drugs confirmed to be causally related to the ADR and plausible ADRs are drugs that have a high likelihood of being causative. The combined signaling system demonstrated a statistically significant large improvement over the FAERS in the precision of top-ranked signals (i.e. from 31 % to almost threefold for different evaluation categories). The study concluded with promising initial evidence that exploring FAERS and EHR data in the scope of replicated signaling can improve the accuracy of signal detection in specific cases. Vilar et al conducted two studies of re-ranking the ADR signals mined from observational health databases. One was based on a single EHR system and the other was based on a large-scale claims database using 2D structure similarity for enrichment analysis (Vilar, Harpaz et al. 2011, Vilar, Ryan et al. 2014). However, the above studies used a single data resource to generate ADR signals and then independently used another resource for validation or enrichment analysis. Harpaz et al. proposed a Bayes model to computationally combine ADR signals from a disparate SRS of about 5 million adverse event reports collected by the FDA and from healthcare data corresponding to about 46 million patients from a healthcare claims database, and the performance was measured based on a reference standard of 4 ADRs and 399 test cases provided by OMOP (Harpaz, DuMouchel et al. 2013). The metrics used were the area under receiver operation characteristic curve (AUC) and partial AUC. Results

demonstrated that the proposed method led to a statistically significant and substantial improvement in signal detection accuracy, averaging 40% over the use of each source independently, and an area under the ROC curve of 0.87. Another advantage of this method is that the method does not require labeled (training) samples whose availability is currently limited. The study of Liu et al. also followed an integrative perspective for ADR detection by utilizing chemical - e.g. compound fingerprints or substructures, biological – e.g. protein targets and pathways, and phenotypic properties of drugs – e.g. indications and other known ADRs (Liu, Wu et al. 2012). This integrative analysis was evaluated based on the prediction of 1,385 known ADRs of 832 approved drugs, through five different analysis methods, namely logistic regression, naive Bayes, K-nearest neighbor, random forest and support vector machine. The detailed data were obtained from public databases, while the evaluation was based on accuracy, precision, and recall, which were determined by the best operating points of the global ROC curve on the basis of the prediction scores for all ADRs. The study indicated that from the three types of information, phenotypic data were the most informative for ADR prediction. However, when biological and phenotypic features were added to the baseline chemical information, the proposed prediction model achieved significant improvements and successfully predicted ADRs associated with the withdrawal of specific drugs.

Patadia et al evaluated performance of electronic healthcare records and spontaneous reporting data in drug safety signal detection on the basis of ten events with known positive and negative reference sets. Signals were identified when respective statistics exceeded defined thresholds. The results showed that when using all cumulative data, signal detection in SRS data achieved higher specificity and sensitivity than EHR data. However, when data were restricted to time prior to a regulatory action, the appropriate use of healthcare data had an potential for earlier

detection of drug safety signals before healthcare professionals report them to an SRS system (Patadia, Schuemie et al. 2014).

It is believed that one of the next breakthroughs in pharmacovigilance depends on a comprehensive approach that examines ADR-related information from a diverse set of potentially complementing data sources such as SRS, electronic healthcare data, biomedical literature, chemical information and phenotypic information, to detect and validate novel ADRs.

### 2.5 Reference standards used in Pharmacovigilance

A central challenge in ADR detection is the need for publically available and sufficiently large reference standards to properly evaluate the performance characteristics of the data mining algorithms when applied to various data sources. There have been previous attempts to develop reference standards, however, the procedure to generate them was not transparent and systematic, or lacked negative controls. For example, Lindquist et al. evaluated the performance of the BCPNN based on the Martingdale and Physician Desk Reference compendium of drug information(Lindquist, Ståhl et al. 2000). Hochberg et al. selected 27 drugs and classified adverse events based on level of evidence from product labeling and literature review, and used this reference event database to evaluate three algorithms (Hochberg, Hauben et al. 2009). Pharmacovigilance research has become an important topic in the biomedical informatics field. Wang et al. conducted a feasibility study of using NLP, Statistics, and EHRs for the pharmacovigilance and selected seven drugs/drug classes with their 132 known ADRs to evaluate the system (Wang, Hripcsak et al. 2009). Harpaz et al and Li et al evaluated their regression-based ADR detection systems on the basis of three ADRs – rhabdomyolysis,

pancreatitis and QT prolongation (Harpaz, Haerian et al. 2010, Li, Salmasian et al. 2013). The reference standard for the known drugs causing these three ADRs was created using evidence from literature, Micromedex and drug labels, and was classified into two categories – established and plausible (Harpaz, Vilar et al. 2012). LePendu et al evaluated their system using the manually curated reference standard of 28 positive associations and 165 negative associations spanning 78 drugs and 12 different events for single drug–adverse event associations (LePendu, Iyer et al. 2013). Quite a few studies evaluated their systems using the popular database - Side Effect Resource (SIDER), which is a publicly available knowledge base that contains a total of 99,423 drug-ADR pairs regarding 4192 ADRs and 996 drugs (Leaman, Wojtulewicz et al. 2010, Nikfarjam and Gonzalez 2011, Shang, Xu et al. 2014). The information in SIDER is automatically extracted from public documents and package inserts but SIDER does not differentiate carefully the degree of certainty for a drug ADR signal when it is appearing in different sections of the drug label, and therefore some drug ADR pairs could be false positive signals. Other evaluations were performed via comparative analysis with findings from previous studies, for example, Caster et al compared the ADR signals generated by Lasso Logistic Regression (LLR) with the ones produced by the routine method information component (IC) used in the Vigibase and found that LLR was able to detect some established drug safety issues earlier than the IC (Caster 2007). Xu et al compared ADR signals detected by mining literature with the ones mined from FAERS (Xu and Wang 2014).

Recent efforts made by the EU-ADR projects and OMOP have made substantial progress in developing reference standards. The EU-ADR projects constructed a reference standard for ten top-ranked events judged as important in pharmacovigilance. A stepwise approach was employed to classify drug-ADR pairs to positive or negative test cases based on MEDLINE-

indexed publications, drug product labels, spontaneous reports made to the WHO's pharmacovigilance database, and expert opinion, resulting in 44 positive and 50 negative test cases, with up to 5 positive and negative controls for 10 ADRs (Coloma, Avillach et al. 2013). In its initial experiments, OMOP constructed a reference standard of 53 drug-ADR pairs which were classified as 9 positive test cases and 44 negative test cases on the basis of product labeling and expert consensus (Ryan, Madigan et al. 2012, Ryan, Schuemie et al. 2013). Later on, they selected four ADRs and classified drugs associated with these ADRs on the basis of evidence from product labeling, systematic review of the literature and a textbook about drug-induced diseases resulting in 399 test cases – 165 are positive cases and 234 are negative controls(Ryan, Schuemie et al. 2013).

The resulting reference standards are by no means definitive, however, and should be seen as dynamic. As knowledge on drug safety evolves over time and new issues in drug safety arise, these reference standards should be re-evaluated and expanded. Therefore, the temporal information is essential about when a true positive drug-ADR signal becomes known or up to when there is no supporting evidence about a drug causing an ADR. Harpaz designed a time-index reference standard, which was systematically curated from drug labeling revisions, such as new warnings, which were issued and communicated by the US Food and Drug Administration in 2013. The reference standard includes 62 positive test cases and 75 negative controls, and covers 44 drugs and 38 events(Harpaz, Odgers et al. 2014). However, the date of revising labeling, are unlikely to truly represent the time of first detection of a new safety signal. For example, Niu et al claimed that the use of data mining with the Vaccine Adverse Event Reporting System (VAERS), the US surveillance system for monitoring vaccine safety, had detected a signal for intussusception earlier than approved drug label (Niu, Erwin et al. 2001). In

fact, the first indication of a potential signal actually occurred prior to approval and was described in the original package insert before it became publicly available. The common evaluation metrics of evaluation are recall, precision, F-score, AUC and partial AUC.

## 2.6 Related Work

### 2.6.1 Natural language processing

The key challenge in using clinical information for pharmacovigilance is that they are represented in free-text. With the emergence of high throughput technologies, natural language processing (NLP) has been applied in biomedicine. A typical procedure to process the free-text clinical notes comprised several NLP subtasks, including named entity recognition, negation detection and relation extraction. A brief description of these tasks is provided by Friedman and Elhadad (Friedman and Elhadad 2014)and Nadkarni (Nadkarni, Ohno-Machado et al. 2011) et al. The commonly used systems in biomedical domain include MetaMap, MedLEE, BioMedLEE and MGrep (Aronson 2001, Chen and Friedman 2004, Friedman, Shagina et al. 2004, Jonquet, Shah et al. 2009). Medical Language Extraction and Encoding (MedLEE) is a natural language processing system that has been used to extract and encode information in clinical narratives for a large number of different applications and studies. For a given report, MedLEE produces a set of findings, such as problem, procedure, device, and medication, along with associated modifiers, such as certainty, degree, status, body location, and section.

NLP was initially proposed to be applied for the active computerized pharmacovigilance by Wang et al (Wang, Hripcsak et al. 2009). They demonstrated that the framework based on NLP, EHR and statistics could potentially unveil drug safety profiles throughout their entire market life. Haerian et al applied the NLP and a knowledgebase to exclude cases in which the patient's

disease was responsible for the event rather than a drug, which is crucial for mining EHR for the detection of ADR (Haerian, Varn et al. 2012). LePendu et al developed a high-throughput NLP tool to transform clinical notes into a feature matrix encoded using medical terminologies, and then used statistical method to detect ADRs (LePendu, Iyer et al. 2013).

### 2.6.2 Biomedical ontologies

The Unified Medical Language System (UMLS) is one of the major resources, which comprise three components: the Metathesaurus, the Semantic Network, and the SPECIALIST Lexicon. The UMLS Metathesaurus is a compendium of over 150 controlled vocabularies or ontologies containing 3 million biomedical concepts that are associated with synonyms, semantic groups and relationships between two concepts (Bodenreider 2004). In addition, the UMLS uses the concept unique identifier (CUI) to link terms with the same meaning together. Among all the contributing sources are two vocabularies utilized commonly in the NLP task for this dissertation. RxNorm (Liu, Ma et al. 2005) is an initiative for creating standard names for clinical drugs, and defining several types of relationships between concepts that are related to generic classes and trade names of drugs, such as *tradename_of* and *has_tradename,* which are used to map all trade name to their generic names (http://www.nlm.nih.gov/research/umls/ , Liu S 2005, Chen, Hripcsak et al. 2008). The UMLS Semantic Network provides a semantic categorization of the UMLS concepts and includes a set of 135 semantic types such as *Disease or symptom* (T047) and *Pharmacologic Substance* (T121), as well as semantic relations defining relations between these types.

Search tool for interactions of chemicals (STITCH) integrates information about interactions from metabolic pathways, crystal structures, binding experiments and drug–target relationships.

STITCH maintains synonym lists for chemicals, and relationships between drugs and their chemical compounds. (Kuhn, von Mering et al. 2008) For example, quinapril hydrochloride and Hemokvin are mapped to the main ingredient quinapril. STITCH was used to link drug brand names to their chemical compound names.

# CHAPTER 3 A method for controlling complex confounding effects in the detection of adverse drug reactions using electronic health records

## 3.1 Introduction

EHRs contain comprehensive patient information collected during routine practice (Cox, Martin et al. 2009). Unlike spontaneous reporting systems, they are not subjective regarding ADRs. However, EHRs introduce other challenges. First, most of the information is buried in narrative clinical notes, and is inaccessible for automated applications. This can be addressed by using natural language processing (NLP) systems, which encode narrative clinical notes (Meystre, Savova et al. 2008, Savova, Masanz et al. 2010, Xu, Stenner et al. 2010). Second, the vast amount of clinical narrative information in the EHR exacerbates the problem of confounding by introducing many conditions. Third, records usually mention the patient's medications, symptoms, diseases, and procedures individually without mentioning their relationships. Therefore, statistical methods are needed to obtain associations, which do not denote relationships. For example, a statistical association between a medication and a condition may be a treatment, an ADR, or an indirect association stemming from another event, for example, a confounder (Cao, Hripcsak et al. 2007, Wang, Hripcsak et al. 2009). Since ADRs occur rarely, most associations are due to confounding. For instance, when certain serious ADRs were identified using abnormal laboratory signals (ALS), 70% were not drug-related, but corresponded to spurious associations between drugs and the adverse events (Ramirez, Carcas et al. 2009). ADR signals detected in the EHR are likely to be confounded by co-medication, by

indication, by comorbidity, or any combination of the three. Confounding by co-medication occurs when two or more medications are frequently prescribed together, but only one causes the ADR of interest. For example, Rosinex causes nausea, but because Rosinex and Ganclex are frequently prescribed together, a spurious association between Ganclex and nausea may also occur (Hauben, Madigan et al. 2005). Confounding by indication occurs when medications are prescribed to treat symptoms or manifestations of an ADR before the ADR is diagnosed. For example, the medication fentanyl may be prescribed for patients who have severe pain before the diagnosis of the condition responsible for the pain. Confounding by comorbidity occurs when an ADR is associated with the disease which the medication is used to treat. For example, Naltrexone may be associated with pancreatitis because it treats alcoholism, which often leads to pancreatitis. In this study, we focus on eliminating confounding by co-morbidity.

To ascertain a causal relationship between a drug and an ADR, confounders need to be identified and removed from the observed marginal associations. A marginal association is a relationship between two variables in the marginal table, and can be used to test for marginal independence between two variables while ignoring the third. Removing confounding effect is critical for observational studies, where the data are collected without randomization or strict inclusion/exclusion criteria (Greenland and Morgenstern 2001, Brookhart, Stürmer et al. 2010). A study conducted by Harpaz *et al* selected potential confounders which were highly associated with the outcome ADR and then determined whether an association between a medication and an ADR existed based on changes in association strengths with and without the confounders.(Harpaz, Haerian et al. 2010) These identified confounders are actually more similar to risk factors (RFs) for an ADR (hereafter Harpaz's method is referred to as RF). The propensity score method (PSM) also controls for confounding, and has been applied to health

claims databases for drug effectiveness comparative studies (Schneeweiss and Avorn 2005, Schneeweiss, Rassen et al. 2009, Brookhart, Stürmer et al. 2010) and ADR detection. (Caster 2007, Caster, Norén et al. 2010, Tatonetti, Ye et al. 2012) The PSM estimates each patient's probability of the exposure of medication, which it uses as a surrogate to mitigate confounding. The RF method identifies the confounders only by their associations with the ADR, while the PSM selects confounders based only on their associations with the medication. In addition to that, PSM selects potential confounders on an individual basis that are often correlated with each other. However, some conditions no longer confound the drug-AE association in the presence of other conditions. Including these unnecessary conditions in the analysis leads to increased uncertainty and decreased statistic power. The algorithm we propose takes both types of associations into account, which helps avoid detecting inappropriate confounders. We apply our method to two serious ADRs, rhabdomyolysis and pancreatitis, to study performance, but it is generalizable and can be used to detect other ADRs.

## 3.2 Method

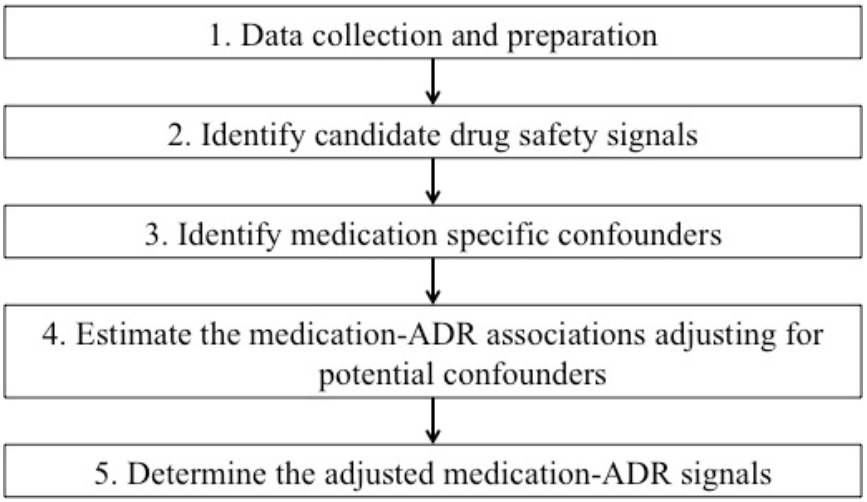
### 3.2.1 Study setting

The study was conducted at Columbia University Medical Center/New York Presbyterian Hospital (CUMC/NYPH), after Institutional Review Board approval. EHR data consisted of retrospective narrative outpatient visits, admission notes, discharge summaries, and structured medication orders and laboratory results from 2004 to 2010. Narrative reports and structured medication orders were used to obtain the patients' medical conditions and medications, and laboratory data was used to detect ADR occurrences.

### 3.2.2 Methodological Framework

Figure 3.1 is an overview of the methodology, which consists of 5 steps: 1) collecting the appropriate EHRs and performing NLP of the narrative notes to obtain structured coded data; 2) creating each ADR case group, generating the 2×2 contingency tables, and identifying initial candidate drug safety signals; 3) identifying potential confounders; 4) estimating medication-ADR associations while adjusting for confounders; and 5) determining medication-ADR signals.

Figure 3.1 Methodological Framework




#### 3.2.2.1 Data collection and preparation

An NLP system, MedLEE, was used to structure and encode the narrative notes. (Friedman, Shagina et al. 2004) MedLEE identified medical concepts, such as medications, diseases and symptoms, and mapped the concepts to the Unified Medical Language System (UMLS) concept unique identifiers (CUIs) to standardize them. (Bodenreider 2004). MedLEE also identified

modifiers of the medical concepts, such as time and negation. By using them, events that were not experienced by the patient or that occurred in the past were excluded.(Chapman, Bridewell et al. 2001) For example, *chest pain* in the sentence "The patient had 3 admissions in the past for chest pain", was excluded as a current problem. Medication names were normalized to their generic names. As mentioned in Chapter 2, the UMLS Metathesaurus includes relationships among concepts drawn from its various source terminologies, and the hierarchic relations provide a basis for normalizing drug brand names to generics and linking specific drugs to drug classes. We retrieved all the "isa," "inverse_isa," "has_tradename," and "tradename_of" relations of each extracted drug concept to create the hierarchy, and also used "has_ingredient" and "has_active_ingredient" relations to help determine whether a concept is a drug or a drug class. For example, the trade name Lipitor (UMLS id entifier C0593906) was normalized to the generic atorvastatin (UMLS identifier C0286651). Finally, we captured temporal information corresponding to dates of the laboratory tests, dates of admission and discharge for inpatients, and dates of office visit for outpatients.

**3.2.2.2 Identify candidate drug safety signals**

The two ADR groups were identified based on abnormal laboratory tests. Rhabdomyolysis was based on a serum CK >= 1000 U/L, (and pancreatitis was based on an amylase >= 300 U/L or lipase >= 120 U/L. The control groups for each ADR consisted of patients in the same population without the particular ADR. We analyzed associations of ADRs by considering medications that were mentioned *before* the ADR occurred as the exposure should always precede the ADR. We utilized two criteria to select medications in the case group: 1) medications mentioned in a clinical note were included if the note was written before the initial date of the abnormal lab signal (ALS), or 2) only medications mentioned in the sections

*Medications on Admission* or *Current Medications* were included if the note was written during the same admission or office visit corresponding to the date of the first ALS because these sections generally specify medications taken prior to the ADR. In contrast, all the medications for the control patients were collected. Subsequently, we constructed 2×2 contingency tables for each medication-ADR pair, as shown in Table 3.1.

Table 3.1 Two by two contingency table

| | ADR (Present of outcome) | No ADR (Absence of outcome) | |
|---|---|---|---|
| Medication (Exposure) | a | b | (a+b) |
| No medication (No exposure) | c | d | (c+d) |
| | (a+c) | (b+d) | (all patients) |

Using formula 3.1 we calculated the Odds Ratio (OR) for each contingency table to obtain an initial set of drugs associated with the ADR. An OR >1 indicates that the chance for developing an ADR is higher for those who took the medication than who did not. We used the Fisher's exact test(Upton 1992) to test whether the ORs were significantly larger than 1, and ranked the resulting p-values from smallest to largest. We selected the top K drugs using a family-wise False Discovery Rate (FDR) (Benjamini and Hochberg 1995) controlled at 5%.

Equation 3.1 Odds Ratio

$$OR_{ADR,Rx} = \frac{Odds(ADR = 1|Rx = 1)}{Odds(ADR = 1|Rx = 0)} \; where \; Odds(X) = \frac{Pr(x)}{1 - Pr(x)}$$

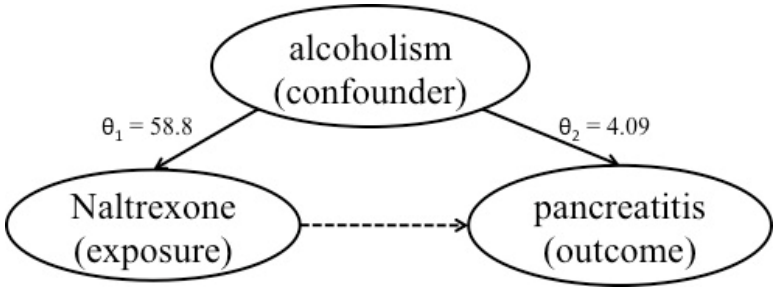
*Rx* represents the drug of interest

**3.2.2.3 Identify confounders for specific medications**

Potential confounders included diseases and symptoms of individual patients. We calculated the OR of each condition with the drug ($\theta_1$), and with the ADR ($\theta_2$), and identified a condition as a

confounder for the drug-ADR association if: (1) both $\theta_1 > 1$ and $\theta_2 > 1$, and (2) $\ln(\theta_1 \times \theta_2) > 0.2$. The rationale is that a confounder could falsely amplify the ADR signal if and only if it is positively associated with both the drug and the ADR, and the associations are strong. For example, as shown in Figure 3.2, alcoholism was positively associated with both Naltrexone ($\theta_1 = 58.8$), and pancreatitis ($\theta_2 = 4.09$), and the associations were strong ($\ln(\theta_1 \times \theta_2) = 5.74$). Therefore, alcoholism was considered a potential confounder for Naltrexone-pancreatitis.

Figure 3.2 Example of confounding by comorbidity




**3.2.2.4 Estimate the medication-ADR associations adjusting for potential confounders**

We fit the logistic regression model shown in formula 3.2 to re-evaluate the drug-ADR association while adjusting for the identified confounders simultaneously.

Equation 3.2 Logistical regression model

$$logit\{prob(ADR = 1)\} = \alpha + \beta Rx + \sum_{i=1}^{M} \gamma_i C_i$$

Rx represents the medication of interest

β is the effect of the medication associated with the ADR after adjusting for all the Cis

$\gamma_i$ is the effect of the i-th confounder $C$ concerning the ADR

In last step, the potential confounders $C_i$ were identified on an individual basis, and were often correlated with each other. Hence some conditions no longer confounded the drug-AE association in the presence of other conditions. Including irrelevant items could inflate the estimation variability and undermine the statistical power for detecting ADR associations. To address such over-controlling, we incorporated a Least Absolute Shrinkage and Selection Operator (LASSO) type regularization into the estimation of the model which automatically selected the significant $C_i$'s (Tibshirani 1996, Zou and Hastie 2005). The LASSO involves a turning parameter $\lambda$, which controls the penalty on the model complexity. We selected an optimal $\lambda$ by ten-fold cross-validation.

To relieve the computational burden, we included the conditions into formula 3.2 in groups instead of all at once. Specifically, we ranked the $C_i$'s by the strength of their association with the ADR ($\theta_2$). Instead of including all the $C_i$'s at once, we only included the top 500 confounders, and then used LASSO to eliminate the insignificant conditions. We repeated this procedure by iteratively adding the next 500 confounders. The method stopped and the drug-ADR association was rejected if after adding confounders, there was no association between medication and ADR. However, if after adding all confounders, the association still existed, this was considered a possible ADR signal.

**3.2.2.5 Determine the adjusted medication-ADR signals**

For each drug-ADR association, we tested the null hypothesis $\beta = 0$ using the Wald test (Gourieroux, Holly et al. 1982). If $\beta = 0$ was accepted, it implied that the observed marginal drug-ADR association was due to the existing confounding conditions; otherwise, the medication was considered to be associated with the ADR even after adjusting for the confounders.

### 3.2.3 Evaluation Design

#### 3.2.3.1 Reference standard

The reference standard consisted of drugs implicated in causing rhabdomyolysis and pancreatitis. It was constructed independently by a pharmacological expert using Micromedex, literature reviews and published reports, and ADRs listed in the Medi-Span adverse drug effects databases, and is described in more detail in another paper (Harpaz, Vilar et al. 2013).

#### 3.2.3.2 Comparisons

Four methods were compared with ours in this study: 1) a baseline method, which only used steps 1 and 2 of the proposed method where confounding was not considered 2) a knowledge-based method where a knowledgebase, developed by clinical experts containing comprehensive non-drug related risk factors for rhabdomyolysis and pancreatitis was applied to exclude patients with predisposing conditions, which eliminated confounders from the population regardless of medication exposures. The rhabdomyolysis knowledgebase was previously established and is in the supplemental data of Haerian's publication,(Haerian, Varn et al. 2012) and the one for pancreatitis is available in Table 3.2. After excluding patients with underlying conditions for developing ADRs, we performed step 2 of the proposed method. 3) the RF method proposed by Harpaz *et al*. was utilized where the shrinkage parameter was selected based on a conjecture that a size of between 20 and 40 conditions was reasonable, however, we used cross-validation to

select the shrinkage parameter since this was more reproducible. 4) the PSM proposed by Tatonetti *et al* was replicated, except that for each medication, we only used the top 200 associated conditions, based on their phi coefficients, to generate the propensity score for each patient.

Table 3.2 Medical conditions that were found to be risk factors for pancreatitis

| UMLS ID | UMLS PREFERRED TERM | UMLS ID | UMLS PREFERRED TERM | UMLS ID | UMLS PREFERRED TERM |
|---|---|---|---|---|---|
| C0085762 | alcohol abuse | C0008340 | choledochal cyst | C0023891 | liver cirrhosis, alcoholic |
| C0001957 | alcohol withdrawal delirium | C0701818 | Choledocholithiasis | C0346647 | malignant neoplasm of pancreas |
| C0236663 | alcohol withdrawal syndrome | C0008350 | Cholelithiasis | C0877425 | mass of pancreas |
| C0156076 | alcoholic gastritis | C0008370 | Cholestasis | C0333027 | Microlithiasis |
| C0001973 | alcoholic intoxication, chronic | C0009438 | common bile duct calculi | C0085407 | Microsporidiosis |
| C0267931 | bile duct cysts | C1397941 | gallbladder distension | C0008313 | cholangitis, sclerosing |
| C0005411 | biliary atresia | C0860209 | gallbladder sludge | C0008320 | cholecystectomy procedure |
| C0242216 | biliary calculi | C0744257 | gallbladder wall thickening | C0008325 | Cholecystitis |
| C0151824 | biliary colic | C0521614 | gallstone pancreatitis | C0149520 | cholecystitis, acute |
| C0282074 | biliary sludge | C0019187 | hepatitis, alcoholic | C0947622 | Cholecystolithiasis |
| C0597984 | biliary stricture | C0020437 | Hypercalcemia | C0026780 | Mumps |
| C1167663 | Biloma | C0020502 | Hyperparathyroidism | C0400976 | obliterative cholangitis |
| C0206698 | Cholangiocarcinoma | C0020557 | Hypertriglyceridemia | C0747181 | pancreas head mass |
| C0008311 | Cholangitis | C0022354 | jaundice, obstructive | C0235974 | pancreatic carcinoma |
| C0030283 | pancreatic cyst | C0030297 | pancreatic neoplasm | C0566602 | primary sclerosing cholangitis |
| C0267919 | primary cholangitis | C0030299 | pancreatic pseudocyst | C0149783 | steroid therapy |

## 3.3 Results

### 3.3.1 Data collection and cohort characteristics

Data was collected for 264,155 patients accounting for 6,221 unique generic drugs and 32,122 unique medical conditions. The characteristics of patients who had rhabdomyolysis and pancreatitis are shown in Table 3.3. There were more men than women, and more African-Americans than other ethnic groups developing rhabdomyolysis as expected because baseline CK levels are higher in men than in women, and higher in African-Americans than in the other groups (Neal, Ferdinand et al. 2009). There were almost equal numbers of men and women, and no ethnic predisposition for pancreatitis. There is no evidence that ethnicity or age affect the chance of developing pancreatitis (Santhi Swaroop Vege).

Table 3.3 Demography of patient population

| Variable | Unique Patients | Rhabdomyolysis | Pancreatitis |
|---|---|---|---|
| N | 264,155 | 3,670 | 6,294 |
| Mean Age (±SD) | 50.9(±23.9) | 57.6(±21.8) | 57.9(±22) |
| Sex (Male) | 42.5% | 68.2% | 50.3% |
| Race (% of group) | | | |
| White | 27.7% | 26.1% | 26.2% |
| Hispanic | 30% | 22.4% | 29.4% |
| Black | 14% | 23.8% | 19.3% |
| Asian | 2% | 2.2% | 2% |
| Other/Undocumented | 26.3% | 25.5% | 23.1% |

### 3.3.2 Reference Standard

Table 3.4 shows statistics and examples of the reference standard.

Table 3.4 The statistics and examples of reference standard

| | Rhabdomyolysis | Pancreatitis |
|---|---|---|
| Total # of drugs | 618 | 436 |
| Examples | acetaminophen, simvastatin, candesartan, iotrolan | amiodarone, omeprazole, meloxicam, zidovudine |

### 3.3.3 Statistics of detected drug-safety signals

True positive signals signify that the signals are in the reference whereas false positive signals signify that those signals are not. Precision is measured as the ratio of true positive signals divided by the sum of true positive and false positive signals. Table 3.5 shows precision for the five methods. Among them, the proposed method performed significantly better than the other four methods for rhabdomyolysis, with a precision of 83.3% compared to 72.7% for PSM, 50% for RF, 58% for knowledge-based method and 38.7% for crude marginal association. For pancreatitis, the proposed method demonstrated similar precision compared with the PSM, as depicted by a precision of 60.8% and 66.2% respectively. The performance of the RF method was comparable to the knowledge-based method, and was worse than the PSM and the proposed methods. The knowledge-based method was significantly better than the baseline method, demonstrating that medical knowledge is effective in identifying confounders, but not as effective as the PSM and proposed models. The number of signals retrieved by each of the five methods is shown in Table 3.6. It is also apparent that the PSM had higher recall than the

proposed methods in terms of more signals detected. The upper bound of the recall for PSM and the proposed model were 0.15 and 0.02 for rhabdomyolysis respectively, and were 0.21 and 0.07 for pancreatitis correspondingly.

Table 3.5 The precision of five methods

| ADR | Crude marginal association (does not deal with confounder) | Knowledge-based method | RF method (only considers conditions for developing ADR) | Propensity score method (only considers conditions for prescribing medications) | Proposed method (considers conditions both to prescribe medications and develop ADR) |
|---|---|---|---|---|---|
| Rhabdomyolysis | 38.7% [33.5%,43.7%] | 58.0% [48.3%,67.7%] | 50.0% [10.0%,90.0%] | 72.7% [65.0%,80.4%] | 83.3% [62.2%,100%] |
| Pancreatitis | 27.7% [24.3%,31.1%] | 32.8% [28.5%,37.2%] | 42.9% [6.2%,79.5%] | 66.2% [58.4%,74.0%] | 60.8% * [47.4%,74.2%] |

The number in the brackets is the confidence interval (CI) for the precision (p)

$CI = p \pm 1.96 * \sqrt{p * \frac{1-p}{n}}$, n is the number of signals retrieved by a method

* This precision can be improved to 70.5% [57.0%, 83.9%] by removing medications treating symptoms of pancreatitis

Table 3.6 The number of signals retrieved by five methods

| ADR | Crude marginal association (does not deal with confounders) | Knowledge-based method | RF method (only considers conditions for developing ADR) | Propensity score method (only considers conditions for prescribing medications) | Proposed method (considers conditions both to prescribe medications and develop ADR) |
|---|---|---|---|---|---|
| Rhabdomyolysis | 364 | 100 | 6 | 128 | 12 |
| Pancreatitis | 666 | 437 | 7 | 142 | 51 |

Table 3.7 lists the true and false positive signals obtained by the proposed method for rhabdomyolysis and pancreatitis. The false positive signals could be classified as due to 1) co-medication confounding, 2) indication confounding, 3) comorbidity confounding, and 4) possible true signals not in the reference standard.

Table 3.7 ADR signals detected by the regression-based method and compared with reference standard

| | | Rhabdomyolysis | Pancreatitis |
|---|---|---|---|
| TP | established | gemfibrozil, olanzapine, atorvastatin | aluminum hydroxide, calcitriol, didanosine, furosemide, pentamidine, propofol, sulfamethoxazole, trimethoprim, lisinopril, stavudine, folate, lansoprazole, lamivudine, caspofungin, omeprazole, nelfinavir mesylate, imatinib mesylate |
| | plausible | aspirin, lorazepam, lisinopril, sulfamethoxazole, zidovudine, sirolimus, labetalol | ergocalciferol, famotidine, fluconazole, gemfibrozil, nadolol, prednisone, sodium chloride, ondansetron, pantoprazole, mycophenolate mofetil, levofloxacin, atorvastatin, rabeprazole, esomeprazole, |
| FP | 1 | calcium acetate, mycophenolate mofetil | NA |
| | 2 | NA | clonidine, fentanyl, meperidine, metoclopramide, norepinephrine, nystatin, simethicone, vancomycin, sodium acetate, calcium acetate |
| | 3 | NA | insulin, nph insulin, ursodeoxycholate, ursodiol, midazolam, lorazepam |
| | 4 | NA | levodopa, sildenafil citrate, lepirudin, sevelamer carbonate |

TP: true positive; FP: false positive; False positive signals are likely due to confounding by 1 co-medication; 2 indication; 3 co-morbidity, and 4 possible true signals not in the reference standard. NA: not applicable

## 3.4 Discussion

Our results demonstrate that the proposed method is effective for dealing with confounders from EHR reports, and either outperforms or has similar performance as the four other comparators.

### 3.4.1 Qualitative Analysis of Results

#### 3.4.1.1 False positive signals

Only two false positive signals were obtained for rhabdomyolysis likely due to confounding by co-medication. For example, calcium acetate treats patients who have transplants or end stage renal disease, and consequently are on multiple drug regimens, such as prednisone and tacrolimus, both of which are known to cause rhabdomyolysis. Our method currently does not handle confounding by co-medication, but will address it in future work.

Among the false positive signals for pancreatitis, six were likely due to confounding by co-morbidity. For example, ursodeoxycholate and ursodiol are used to treat gallstones, common bile duct calculi, and biliary cirrhosis, which are risk factors for pancreatitis. After controlling for these confounders, the association still existed between pancreatitis and those two medications. This could be due to inherent limitations of EHR documentation, NLP errors produced during data preparation, or using incorrect time sequences in patients with only a single visit.

The other 10 false positive signals, such as fentanyl, were likely due to confounding by indication. According to our criteria, these drugs should have been excluded as the exposures occurred after the ALS. However, for some cases the first measurement for amylase/lipase occurred after the drugs were ordered, which mainly happened because treatment for pancreatitis

was started based on early symptoms before the ALS was obtained, or because of the data characteristics, which is explained below. Such false positives are categorized in Table 3.8.

Sevelamer carbonate, lepirudin, sildenafil citrate and levodopa are four candidates for which physicians could not find confounding or other reasons to relate with pancreatitis. Further investigation of these drugs will be performed in future work.

In this study, we compared results to a reference standard but did not look at individual cases to determine what the actual causes of the ADR were for the individual patients, therefore some true positive signals may be false when applied to patients.

Table 3.8 Error analysis for false positive signals associated with pancreatitis

| Possible relationship with pancreatitis | Medication |
|---|---|
| Treatment for comorbidity of pancreatitis | |
| 1. Treatment for gallstones that can cause pancreatitis | ursodiol, ursodeoxycholate |
| 2. Treatment for stopping alcoholic abuse that can cause pancreatitis | lorazepam, midazolam |
| 3. Treatment for hypertriglyceridemia that can cause acute pancreatitis or pancreatic problem induced diabetes mellitus | insulin, nph insulin |
| Treatment for symptoms of pancreatitis or pancreatitis-induced problems | |
| 4. Treatment for pain associated with pancreatitis | fentanyl, meperidine |
| 5. Treatment for pancreatitis-induced vasodilatory shock | norepinephrine |
| 6. Treatment for pancreatitis-induced infections | nystatin, vancomycin |
| 7. Treatment to reduce pancreatic juice secretion. It could be also used to treat a stress situation in pancreatitis with high catecholamine levels. | clonidine |
| 8. Regulation of sodium and calcium disorders associated to pancreatitis | sodium acetate, calcium acetate |
| 9. Used as an antiemetic in patients with pancreatitis. | metoclopramide |
| 10. Reduction of bloating in patients with pancreatitis. | simethicone |

#### 3.4.1.2 False negative signals

False negative signals signify that the signals were not detected by the method but are in the reference standard. There were two reasons for false negative signals: insufficient data and over-adjusting.

Having a large enough set of patients is critical for detecting ADRs, especially rare cases (Makuch 2006). For example, in order to detect chloroquine-induced rhabdomyolysis (incidence rate between 3% and 5%), at least 100 patients must take this medication (Tisdale and Douglas 2010). However, in our data set, there were only 37 patients on chloroquine. An insufficient number of patients for certain medications seemed to be the primary reason for false negatives.

False negative signals also occurred due to over-adjusting, where the proposed method selected more confounders than it should have. For example, amlodipine, which causes pancreatitis between 1% and 4% of the time, was prescribed to 28,832 unique patients in our data, but the proposed method did not detect this since it adjusted for several superfluous confounders such as cytomegalovirus infection. In the future we will explore considering conditions based on smaller p-values to address this problem.

### 3.4.2 The characteristic of the data set and the inherent nature of the two ADRs

The results showed that the proposed method obtained better precision for rhabdomyolysis than for pancreatitis, which is due both to the characteristic of the data and to the nature of the ADRs. About 42% of the data set we used consists of patients with only a single visit. In such a case, when the ALS is reported, the corresponding clinical note frequently mentions the ADR, which is a diagnosis based on the ALS. In that sense the ALS and ADR are synonymous, and the ADR

is not a confounding condition. For example, a patient with an elevated CK test is likely to have rhabdomyolysis mentioned in their note. Therefore, we eliminated use of the conditions rhabdomyolysis and pancreatitis respectively when they occurred in the note associated with the same hospitalization as the ALS. The strategy worked well for rhabdomyolysis but not for pancreatitis because rhabdomyolysis is typically an acute event. In contrast, pancreatitis could also be chronic, and chronic pancreatitis may lead to an ALS, or a predisposition for acute episodes. Therefore, removing mentions of pancreatitis reduced our method's ability to detect it as a confounder, leading to reduction in precision for detecting the ADR pancreatitis. If the EHR included more instances of multiple visits, we should have been able to differentiate chronic from acute conditions. We subsequently explored the false positive medication signals by allowing pancreatitis to be a confounder if it met the criteria of confounding for category 2 and 3 of Table 3.5, and eight false positive signals were removed which are displayed in Table 3.9, improving the precision of the proposed method from 60.8% to 70.5% [57.0%, 83.9%]. Although we were aware of the problem caused by single visits, we included them in the data set because it was critical to obtain as many medication events as possible. Another difference between the two ADRs is that confounding by indication does not occur for rhabdomyolysis because medications are not used to treat it, but confounding by indication must be handled for pancreatitis since medications are used to treat it.

Table 3.9 The eliminated drug safety signals after the post hoc analysis for the pancreatitis

| The category of false positive signals | Medications |
|---|---|
| 2 | clonidine, meperidine, metoclopramide, nystatin, simethicone, vancomycin |
| 3 | nph insulin, lorazepam |

### 3.4.3 Comparison of methods

Apart from performance, the proposed method has the advantage of generalizability over the knowledge-based method. Generalizability is important because different facilities may have different populations. For example, Ramirez *et al* identified *burn* as a major cause for rhabdomyolysis in their population;(Ramirez, Carcas et al. 2009) in contrast, Haerian *et al* found that *myocardial infarction* was a major cause for elevated Creatine Kinase (CK, the laboratory test for rhabdomyolysis) in their population (Haerian, Varn et al. 2012). Developing knowledge specific to each population requires that expertise and manual review of patient charts to select risk factors, which is costly. In comparison, the proposed method automatically identifies and adjusts for confounders. In addition, the proposed method determines confounders in a data-driven fashion, which allows for finding *proxy* variables for the confounders, whereas the confounders must be predetermined when using knowledge. For example, in the association between aspirin and rhabdomyolysis, our method correctly identified ST-elevation myocardial infarction (STEMI) as a confounder (myocardial infarction also causes elevated CK), but also identified *chest pain* and *increased sweating* as confounders, which are common symptoms of myocardial infarction (MI). Our method was capable of adjusting for the confounding effect of MI using these proxy variables. Similarly, our method listed *agitation* and *confusion* as confounders of the association between lorazepam, which is used to treat cocaine abuse, and rhabdomyolysis. Cocaine abusers usually present with agitation and confusion, and are also associated with elevated CK (Warrian, Halikas et al. 1992).

The proposed method has two advantages over the PSM. First, it has the power to detect drug safety signals when it mixes with the effect of comorbidity on the ADR outcome. For example, sevelamer is uniquely prescribed to patients on dialysis, which predisposes them to pancreatitis.

The PSM eliminates the effect of sevelamer on pancreatitis due to the effect of a variety of kidney problems, while the proposed method adjusts for the appropriate confounders including kidney failure, but retains the effect of sevelamer on pancreatitis. Another advantage is the informative clinical knowledge displayed by the confounders identified by the model. For each medication-ADR pair, the proposed method generates a set of confounders, which describes the effect or non-effect of a medication when taking several conditions into account. These conditions provide informative clinical knowledge useful for further analysis of the data. For example, chronic pancreatitis should have been a qualified confounder, but was missing from the pancreatitis model, as explained above. Therefore, we could re-analyze by including pancreatitis in the model. In contrast, the PSM is a black box and is not capable of providing insight concerning confounders. PSM has higher recall while lower precision than the proposed method. In terms of identifying true positive signals in the upper bound level, PSM identifies 83 and 63 more signals than the proposed method for rhabdomyolysis and pancreatitis respectively. Higher recall is important for some ADR tasks, such as early ADR detection, but higher precision is important for others, such as re-ranking potential signals.

### 3.4.4 Use of EHR narratives

There are several advantages to using EHR narratives for detecting ADR signals. It is possible to obtain more comprehensive and finer grained medical information than the International Classification of Disease, Version 9 (ICD-9) codes assigned for billing purposes (Trifiro, Fourrier-Reglat et al. 2009, Ryan, Madigan et al. 2012). Based on our data, patients had about 46 medical conditions on average per year based on their notes, while they only had about 9 ICD-9 codes on average per year. Moreover, had an ICD-9 code, such as cardiac valve fibrosis.

### 3.4.5 Limitations

One limitation of this study concerns time intervals relevant to ADR detection. Currently, the method retrieved all medications prior to an ADR without considering time windows. For instance, a patient who took a drug in 2004 may have discontinued it in the same year, and may have developed an ADR in 2010. Our method counted this patient in the case group but that time interval may be inappropriate. However, note that one general time window cannot be used for detecting all ADRs as previous studies have shown that the window between first drug exposure and the incidence of drug-induced pancreatitis can range between 1 and 1,000 days depending on the drug (Badalov, Baradarian et al. 2007). Also, we confronted the challenge of inadequate documentation or of an incomplete record of patients' health events.

Second, our method did not deal well with other confounding issues, such as protopathic bias, particularly when patients had only a single visit. Protopathic bias occurs when a drug is prescribed for an early manifestation of a disease that has not yet been diagnosed. We plan on collaborating with researchers at other facilities to collect more longitudinal EHR data, which will allow us to obtain more time information.

Third, we used abnormal lab results as surrogates for determining rhabdomyolysis and pancreatitis, which is common in pharmacovigilance, but an abnormal CK could be due to strenuous exercise and not to rhabdomyolysis, and an increased amylase could be due to an inflamed parotid gland, and not to pancreatitis.

## 3.5 Summary

We proposed a novel data-driven method to control for the problem of confounding when using comprehensive EHR data, and demonstrated that the method achieved either a higher or similar

precision in detecting signals for two serious ADRs rhabdomyolysis and pancreatitis when compared to the four other methods while providing insight into confounders for each specific medication-ADR pair. This method is likely to perform better with a larger patient population with more longitudinal data, can be generalized to detect other ADRs while taking into account either an acute or chronic status, and can be easily adapted to other EHR systems.

# CHAPTER 4 A Study of the Effect of Data Characteristics on Adverse Drug Reaction Detection Algorithms in Pharmacovigilance

## 4.1 Introduction

In the United States, the Food and Drug Administration (FDA) maintains the Adverse Event Reporting System (FAERS), consisting of suspected ADR case reports. Among drugs mentioned in a case report, only one is assigned as *primary suspected drug* and others are assigned as *secondary suspected*, *interactive,* or *concomitant drugs*. Some case reports also provide indications for primary suspected medication and patient demographics. Indications and ADRs are coded using the Medical Dictionary for Regulatory Activities (MedDRA) terminology. For example, the MedDRA term *diabetes mellitus* is encoded as an ADR for the medication diazoxide and as an indication for the medication sitagliptin.

Automated methodology has become a standard tool to discover ADR signals from a collection of case reports (Hauben, Madigan et al. 2005, Harpaz, DuMouchel et al. 2012). Disproportionality analysis (DPA) is the main algorithm to detect ADRs, and quantifies the interestingness of each drug ADR pair in the data (van Puijenbroek, Bate et al. 2002, Zorych, Madigan et al. 2013). The most widely cited measurements include relative reporting ratio (RRR), reporting odds ratio (ROR), the geometric mean of empirical Bayes posterior distribution of the "true" RRR (EBGM) produced by Multi-item Gamma Poisson Shrinker (MGPS) and information component (IC) produced by Bayesian confidence propagation neural network (BCPNN). However, the above methods usually include drugs listed as primarily suspected of having caused an ADR without considering concomitant drugs (Caster, Norén et al. 2010,

Tatonetti, Patrick et al. 2012). This situation may vary and include concomitant medications in addition to primary suspected medications for the disproportionality analysis. Drugs usually listed as primary suspects may be reported more often for well-known ADRs. Additionally, these measures do not consider the impact of other variables, which may adversely affect signal detection. For example, indirect associations between a drug and an ADR may result when a drug is frequently co-reported with another drug that causes the ADR. For example, darunavir, which is not known to cause acute renal failure (ARF), is frequently co-prescribed with tenofovir disoproxil, which does cause ARF. An indirect association may also occur when a drug, such as acarbose, treats a condition leading to an increased risk for developing the ADR. For example, acarbose treats diabetes, which predisposes patients to developing ARF. These situations are known as confounding effects of variables not accounted for in the analysis. Therefore, a method that can adjust or control for confounding provided by co-medications and indications should be valuable for ADR detection.

An adjustment by stratification to mitigate confounding effects was first proposed for pairwise associations, which adjusts for age, gender and reporting trend, in case that a particular drug may have different effects among patients with different ages (DuMouchel 1999). Nonetheless, this is not feasible for moderate to large numbers of potential confounders and is only appropriate in the absence of effect modification, which occurs when the effect of two variables are dependent (DuMouchel 1999). Shrinkage regression, such as Bayesian logistic regression and L1 regularization, has been proposed to deal with a large number of potential confounders such as confounding by co-reported medicines, and been demonstrated its success (Caster 2007, Caster, Norén et al. 2010). Another method, named propensity scores (PS), uses co-medications and indications to estimate each patient's probability for the exposure of medication, and then

matches against the case control group based on these scores to mitigate the confounding effect. Tatonetti et al. used the PS method to detect drug-drug interactions in FAERS data, and demonstrated that the method mitigated the confounding effect by showing that the distribution of propensity score for prescribing a medication were balanced across different age and gender groups (Tatonetti, Patrick et al. 2012). Unlike other methods, this PS method considered not only primary suspected medications but also concomitant medications and indications.

Currently, there is no such a study systematically examining the accurateness of primary suspected information delivered by reporters, and evaluating the effect on the ADR detection when considering primary suspected information only or overall information.

This paper proposed a method of two-step LASSO regression to leverages primary suspected medications with concomitant medications and indications. We studied the above two questions by applying the proposed method and three other methods, which are frequency-based method, ROR and GPS, to two data sets: one is on the basis of primary suspected medications (indications for the proposed method), and the other is on the basis of primary suspected, secondary suspected and concomitant medications (indications for the proposed method).

A reference standard was introduced as a benchmark against which four methods can be measured and consequently the studied questions can be answered. The reference standard is consisted of four ADRs: acute renal failure (ARF), acute liver injury (ALI), acute myocardial infarction (AMI) and gastrointestinal bleeding (GI bleeding), and provided by Observational Medical Outcomes Partnership (OMOP) group (Stang, Ryan et al. 2010)(Stang, Ryan et al. 2010)(Stang, Ryan et al. 2010). The area under the receiver operating characteristic (ROC) curve (AUC) are the evaluation metric in this study, which are frequently used to evaluate accuracy of a statistical model (Manning, Raghavan et al. 2008), (DeLong, DeLong et al. 1988).

## 4.2 Method

### 4.2.1 Study Setting

This study used the publicly available FAERS database from 2004 to 2010. Drugs are entered into a report using free text, which can be brand or generic names, while suspected ADRs are coded using MedDRA terms. In order to gain statistical power, we normalize drug names to their chemical compounds using the STITCH database, which maintains synonym lists for chemicals, and relationships between drugs and their chemical compounds (Kuhn, von Mering et al. 2008). For example, quinapril hydrochloride and Hemokvin are mapped to the main ingredient quinapril.

Two data sets were created to study the effect of primary suspected information on ADR detection. The first data set was consisted of primary suspected medications and their indications, and the second data set comprised the secondary suspected and concomitant medications in addition to those had been included in data set 1. The confounding information is represented by indications in data set 1 and signified by all medications and indications in data set 2.

### 4.2.1 Methodology Framework

The proposed method was based on a previously published work conducted by our group which included identifying confounders for specific medications using marginal odds ratios (ORs) and estimating the drug-ADR associations using a least absolute shrinkage and selection operator (LASSO) type regularization (Li, Salmasian et al. 2013). Results showed that the method outperformed the high-dimensional propensity score method, but the resulting false positive rates still exceeded the nominal level (Li, Salmasian et al. 2013). Therefore we revised the method in

two aspects: (1) in the previously work, we only considered the potential confounders that were significantly and positively associated with both the ADR and the medication. We now expanded this list to include medical conditions that were significantly associated with the ADR and medication in either a positive or negative direction. The rationale is that negatively associated conditions could also bias the strength of association. (2) Standard LASSO implicitly assumes a sparse structure in the covariates, and hence tends to select insufficient confounders in high-dimensional regression, which in turn leads to inflated false positive rate. We adopted a two step LASSOs (Belloni, Chernozhukov et al. 2013) for a better control of the false positive rate. In the first step, shown in formula 4.1, standard LASSO is applied to select a set of potential confounders associated with the ADR, denoted by $S_1$; In the second step, shown in formula 4.2, LASSO type regression is used again to select medical conditions that are highly associated with the drug use, and denote them as $S_2$. In both steps, we used 5-fold cross-validation to select LASSO penalties. Finally, we estimate the conditional association between the ADR and drug adjusting for all the confounders in ( $S_1 \cup S_2$). We then use one-sided p-values of the adjusted log odds ratios (log ORs) in the last step as the signal scores, shown in formula 4.3.

Equation 4.1 The first step of two-step LASSOs

$$logit(prob(ADR = 1)) = \alpha^{(1)} + \beta^{(1)}Rx + \sum_{i \in M} \gamma_i^{(1)} C_i$$

Equation 4.2 The second step of two-step LASSOs

$$\mathrm{E}[\omega Rx] = \alpha^{(2)} + \sum_{i \in M} \omega \gamma_i^{(2)} C_i \quad where$$

$$w = \sqrt{prob(ADR = 1|Rx, C_i) * (1 - prob(ADR = 1|Rx, C_i))} \ and \ i \in S_1$$

Equation 4.3 The logistic regression model when controlling for all confounders

$$logit\big(prob(ADR = 1)\big) = \alpha^{(3)} + \beta^{(2)}Rx + \sum_{i \in S_1 \cup S_2} \gamma_i^{(3)} C_i$$

### 4.2.2 Comparators

We used three different methods to compare with the proposed method. The first method, called FREQUENCY, is measured by the number of reports associated with a particular drug-ADR pair and then normalized by the total number of reports corresponding to the same ADR. A higher frequency for a particular drug-ADR pair represents more interestingness. The second is the lower limits of 95% Empirical Bayes Geometric Mean of RRR, called EB05 and the third is the lower 2.5$^{th}$ percentage of ROR distribution, called ROR05 Both the ROR05 and EB05 are DPA methods only using primary suspected information, however, EB05 takes sampling variance into account. ROR05 also represents unadjusted association - not controlling for confounders – of the proposed method.

By comparing performances of different combinations of methods and data sets, we can study the accuracy of primary suspected information described by reporters in terms of whether they are confounded by indications and other medications besides the primary suspected one.

### 4.2.3 Evaluation Metrics

We use ROC and AUC to evaluate performance. An ROC is a graphical plot which illustrates performance of a scoring system as its discrimination threshold is varied (Fawcett 2006). To further compare different scoring systems we reduce ROC performance to a single scalar value representing expected performance by calculating the AUC (Huang and Ling 2005). Both metrics require ranking drug safety signals based on a specific association measurement. Therefore, we generate ranked signals for all the methods for evaluation. We also test the two

sided p-value for the hypothesis of no difference between two AUCs using DeLong's non-parametric approach for correlated ROCs (DeLong, DeLong et al. 1988). In order to make impartial comparison, the score of 0 is assigned to each drug-ADR pair in cases where there are no reports of explicitly corresponding to a specific drug ADR pair. However, if a drug ADR pair has never occurred together in the data set 2, it will be removed from the evaluation.

## 4.3 Results

### 4.3.1 Data characteristics

In total, the accumulated data set from 2004 to 2010 in FAERS had 2,720,634 case reports. The reference standard includes 365 test cases whereas the drug and the ADR are mentioned together at least once. Table 1 shows the number of test cases for each ADR.

Table 4.1 Number of test cases in the reference standard for each ADR

| ADR | Positives | Negatives |
|---|---|---|
| ARF | 23 | 52 |
| ALI | 77 | 33 |
| AMI | 34 | 59 |
| GIB | 24 | 63 |

### 4.3.2 AUCs for different methods and data sets

Table 4.2 AUCs for each combination of ADR, methods and data sets

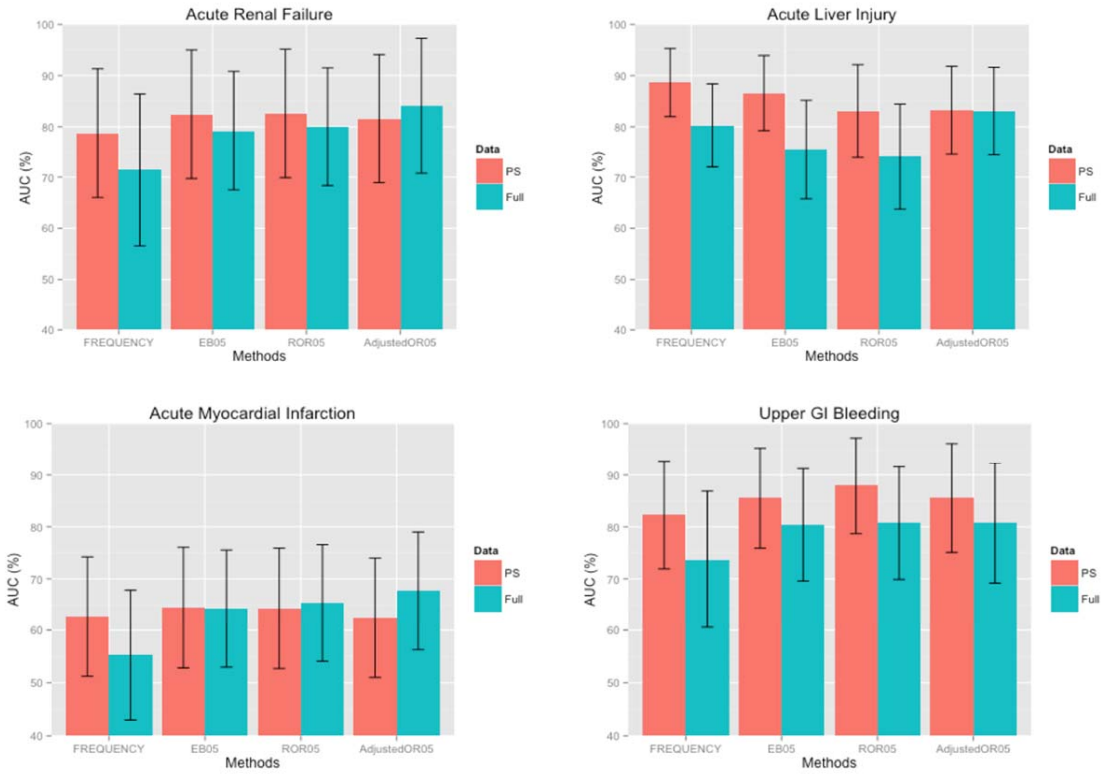
| ADR | Method | PS | Full |
|---|---|---|---|
| AKI | FREQUENCY | 78.68 | 71.49 |
| | GPS | 82.36 | 79.18 |
| | ROR05 | 82.53 | 79.93 |
| | AdjustedOR05 | 81.52* | 84.03 |
| | | | |
| ALI | FREQUENCY | 88.63 | 80.20 |
| | GPS | 86.58 | 75.48 |
| | ROR05 | 83.00 | 74.10 |
| | AdjustedOR05 | 83.16* | 83.00 |
| | | | |
| AMI | FREQUENCY | 62.74 | 55.41 |
| | GPS | 64.46 | 64.26 |
| | ROR05 | 64.31 | 65.35 |
| | AdjustedOR05 | 62.51* | 67.65 |
| | | | |
| GI | FREQUENCY | 82.34 | 73.74 |
| | GPS | 85.58 | 80.42 |
| | ROR05 | 87.96 | 80.75 |
| | AdjustedOR05 | 85.62* | 80.75 |

PS: data consisted of primary suspected medications (indications is used only for AdjustedOR05)

Full: data consisted of primary suspected, secondary suspected and concomitant medications (indications is used only for AdjustedOR05)

* AdjustedOR05 regards indications as potential confounders

Figure 4.1 AUCs of each method based on two different data sets for each ADR



## 4.4 Discussion

### 4.4.1 Quality of primary suspected medication information

Using only primary suspected medications leads to higher AUC performance but it is not significant than using overall medication information except for AdjustedOR05. FREQUENCY achieves fairly good performance compared with other statistical methods, which signifies that reporters are generally accurate when linking the ADR to its causative medication. ROR05 accomplishes better AUC than adjustedOR05 indicating that confounding by indication is less of a problem in data set 1. However, the criteria for constructing the reference standard, such as no statistical evidence – EB05 > 2 (similar to ROR05 when sample size was big) – in FAERS for a negative test case, were correlated with tested methods and therefore could bias the results.

The performance for lower 2.5 percentile of relative reporting ratio (RRR05) is almost identical to ROR05. EB05 is the Bayesian version of RRR05 and has the similar performance with ROR05 expressing that the issue of small sampling variance is not substantial in this study.

All methods are more effective in identifying the other ADRs than AMI. No single method performed consistently better than the other methods for the 4 ADRs based on two data sets.

### 4.4.2 Advantages of the proposed method

AdjustedOR05 attains higher performance than ROR05 when applied to data set 2, which possibly demonstrates the existence of confounding by co-medication but it could also signify that the reference standard is less interrelated with ROR05.

AdjustedOR05 is the best method among overall methods when using the full data set. In addition, it achieves comparable performance with the cutting edge method EB05 on the basis of

primary suspected medications. This is encouraging since AdjustedOR05 carries more information since quite a few drug ADR pairs (128 pairs) were tied to each other at 0 using information of primary suspected medication without sacrificing AUC performance.

### 4.4.3 limitations

One limitation of this study is that it uses a reference standard that is not independent from data and some statistical methods, which may predispose methods using primary suspected information to perform better. Moreover, confounding by unmeasured covariates remains a potential source of misinterpretation that should always be considered in the analysis of outstanding reporting patterns. Another limitation is that case reports in FAERS often contain inaccurate information. For example, some reports mentioned that patients took more than 20 drugs, which may be caused by errors from reporters entering the medication history instead of the medications taken at the time of the report. Finally, we did not deal with duplicate reports, which are known to exist in FAERS and which could falsely lead to a signal.

## 4.5 Conclusion

This study adapts an existing method in a novel way to leverage primary suspected medications with concomitant medications and indications. By comparing performance with three other methods on different data characteristics, we demonstrated that the proposed method generally achieved comparable performance with the state of art GPS method. Methods using primary suspected information generally outperform methods that treat medications equally. However, no single method performed best in detecting all four ADR signals.

# CHAPTER 5 A method to combine analyses from spontaneous reporting systems and observational healthcare data to detect adverse drug reactions

## 5.1 Introduction

Analysis of spontaneous reports of suspected ADRs has traditionally served as a valuable tool in the detection of previously unknown ADRs in post-market surveillance(Bate and Evans 2009, Harpaz, DuMouchel et al. 2012). However, because of well-known limitations of spontaneous reports, such as underreporting and biased reporting, reports (Alvarez-Requejo, Carvajal et al. 1998), electronic healthcare data, such as electronic health records (EHRs) and administrative claims data, are starting to be used to complement the spontaneous reporting system (SRS) (Wang, Hripcsak et al. 2009, Stang, Ryan et al. 2010, Coloma, Schuemie et al. 2011, LePendu, Iyer et al. 2013). However, observational healthcare data has its own limitations such as confounding. Although both SRS and healthcare data represent unique challenges in their use, some researchers believe that they complement each other along several dimensions that may improve pharmacovigilance (Harpaz, DuMouchel et al. 2013, Patadia, Schuemie et al. 2014). Another challenge accompanied with the richness of information for pharmacovigilance practice occurs when these two resources provide conflicting or inconsistent information. Therefore, we propose a methodological framework to integrate analyses generated from the FDA Adverse Drug Event Reporting System (FAERS) and from healthcare data. Harpaz et al's method also combined signals from different sources but imposed the assumptions that the signals generated from each data source be on approximately the same scale, and be log-normally distributed

whereas our method does not impose these assumptions (Harpaz, DuMouchel et al. 2013). As part of the methodological framework, we incorporate a method to deal with confounding effect in NYP/CUMC EHR and the FAERS SRS. We apply the method to four clinically serious ADRs: acute renal failure (ARF), acute liver injury (ALI), acute myocardial infarction (AMI), and upper gastrointestinal bleeding (GIB) (Trifirò, Pariente et al. 2009) with an aim of demonstrating that signal detection performance can be improved by such an integrative strategy. The proposed integrative method is studied using three different experiments aimed at exploring the effect of data size and bias on the method: one where we combine FAERS with a single small-scale EHR database NYP/CUMC, one where we combine FAERS with a large-scale network-based EHR database GE, and one where we combine FAERS with a much larger-scale claims database. We further evaluate the proposed system under the scenarios that the two resources used in combining provide consistent/inconsistent information.

## 5.2 Methodology

### 5.2.1 Data Sources

#### 5.2.1.1 FAERS

The data were extracted from FAERS from 2004 to 2010 encompassing 2.7 million reports, which comprised case reports mainly reported from pharmaceuticals, and to a lesser extent, from healthcare professionals and consumers . We preprocessed and mapped the free-text drug names to their ingredient level specification using the STITCH database(Kuhn, von Mering et al. 2008). The ADRs in FAERS were already coded using MedDRA preferred terms. In this study, we did not utilize the explicit relationships between drugs and ADRs and considered all relationships as co-occurrence information. Consequently, we extended data to all medications mentioned in the case reports including primary suspected, secondary suspected and concomitant, as well as

indications. The signals from FAERS were obtained using the confounding adjustment method, which is presented in the Methods section of Chapter 4.

#### 5.2.1.2 NYP/CUMC EHR

The data, consisting of 0.3 million patients, were extracted from the single-hospital EHR system at NYP/CUMC, after institutional review board approval. The data consisted of retrospective narrative records of inpatient and outpatient visits from 2004 to 2010, including admission notes, discharge summaries, lab tests, structured diagnoses in the form of International Statistical Classification of Diseases, Version 9 (ICD9) codes and structured medication lists. The majority of the data available for this study were from an inpatient population. Narrative reports were used to obtain the patients' medications, and the structured ICD9 diagnosis codes were used to detect ADR events; these codes also served as surrogates of patient characteristics for confounding adjustment analysis. Similar as for FAERS, the signals from the EHR were computed using the confounding adjustment method proposed in this study, which is described in the Methods section of Chapter 4.

#### 5.2.1.3 GE EHR

The EHR database, GE MQIC (Medical Quality Improvement Consortium), represents a longitudinal outpatient population of 11 million patients, and captures certain events in structured form that occur in usual care, including patient problem lists, prescription of medications, and other clinical observations as experienced in the ambulatory care setting. The data were analyzed systematically under OMOP using seven commonly used methods for 399 drug-ADR pairs(Ryan, Stang et al. 2013). The resulting signal scores are reported and publicly available in OMOP. The signal scores for this database were computed using the optimal analytic method for each outcome as follows: self-controlled case series (SCCS) method for

ARF (analysis-ID 1949010), self-control cohort (SCC) method for ALI (analysis-ID 409002), and information component temporal pattern discovery (ICTPD) method for AMI and GIB (analysis-IDs 3016001 and 3034001) (Ryan, Stang et al. 2013).

#### 5.2.1.4 Claims data

In this study, we obtained signal scores associated with the largest claims database, MarketScan Commercial Claims and Encounters (CCAE), which contains information on approximately 46 million patients. Similar to the GE data, CCAE data were extensively analyzed in OMOP for the same drug-ADR pairs with various methods. The signal scores we used for this database were computed by OMOP using the SCC method for ARF, ALI and AMI (Analysis-IDs 404002, 403002 and 408013), and the SCCS method for GIB (Analysis-ID 1931010) (Ryan, Stang et al. 2013).

#### 5.2.1.4 Reference Standard

The reference standard was developed by OMOP. It contains 165 positive and 234 negative controls, i.e., drugs for which there is or is no evidence for corresponding ADRs. This reference set was established by OMOP based on natural language processing (NLP) of structured product labels, systematic search of the scientific literature, and manual validation. The reference standard comprises 181 drugs and four clinically important ADRs: acute renal failure (ARF), acute liver failure (ALI), acute myocardial infarction (AMI), and upper gastrointestinal bleeding (GIB). More details about the reference standard data collection, including drug names, can be found in a previous publication (Ryan, Schuemie et al. 2013).

Other important research conducted by OMOP resulted in establishment of varied ADR definitions, from narrow to broad, for each ADR outcome they studied (Harpaz, DuMouchel et al. 2013, Reich, Ryan et al. 2013). Furthermore, the mapping between ICD-9 codes and

corresponding MedDRA codes for each ADR outcome were also made available by OMOP. We adopted these definitions to identify ADR case groups in NYP/CUMC EHR and in FAERS.

### 5.2.2 Cohort identification

In this study, we used the broad definitions of ICD-9 codes established by OMOP for identifying ADR events in NYP/CUMC HER (Reich, Ryan et al. 2013). The same definitions were also utilized in the GE EHR and the claims database. In addition, we used the corresponding MedDRA codes (as determined by OMOP) for FAERS to identify patients with a particular ADR. Our aim was to ensure that the ADRs are equivalent when using the different databases.

#### 5.2.2.1 FAERS

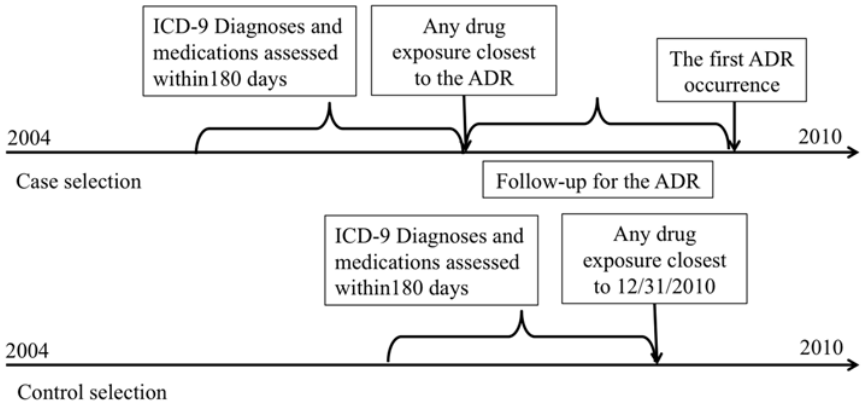
Case reports, which have at least one applicable ADR MedDRA code for an ADR, were identified as a case group, whereas the rest were used as a control group. The indications and all the medications reported in case reports were included as candidate covariates for confounding assessment.

#### 5.2.2.2 NYP/CUMC EHR

The four ADR case groups were identified using their equivalent ICD9 codes. For each ADR, the control group consisted of those patients without the particular ADR. A patient may have multiple records in an EHR and therefore may have experienced an ADR several times, and may have been on and off a particular medication. Only the first occurrence of an ADR was considered and candidate medications were restricted to those that were mentioned before the ADR. If a case patient did not have any medications mentioned before the ADR, or a control patient did not have any medication recorded before 2010, they were excluded from the analysis. We also applied a 180-day window before the latest medication prior to the ADR to retrieve

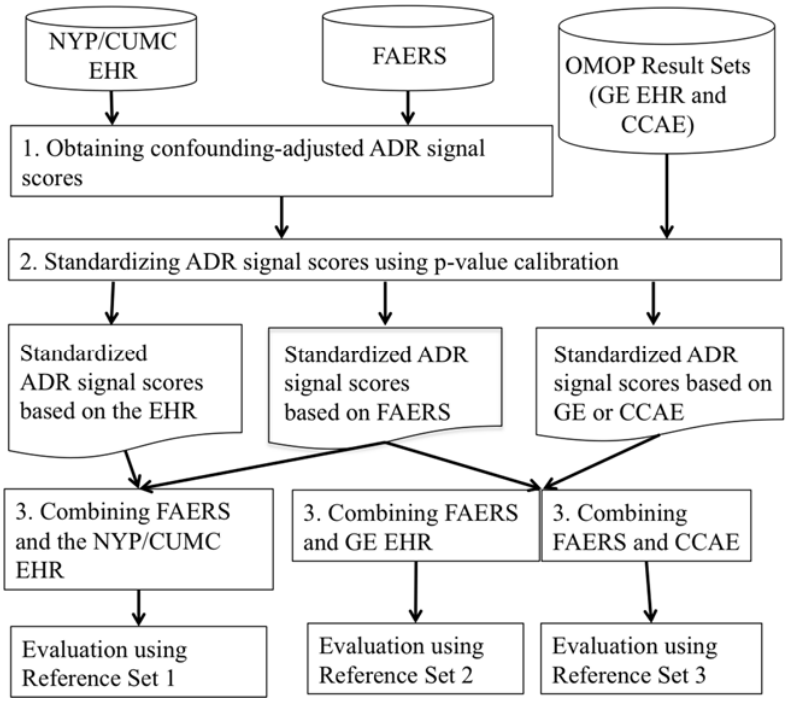
medications and medical conditions (ICD-9 diagnosis codes). We assumed that anything prior to that window are unlikely to be associated with the ADR. For example, a drug taken in 2004 unlikely leads to the development of an ADR in 2010. For the control groups, we used the latest medication record before December 31st, 2010 as the anchor, and retrospectively drew a 180-day window to select medications and ICD-9 diagnoses. Since our patient population was dominated by inpatients with single hospitalization, the individual studying windows in the control groups were evenly distributed from 2004 to 2010. However, the temporality between medications and ADRs could be inaccurate since two types of information occur in the same visit note. Only ICD-9 codes were included as possible confounder candidates. Figure 5.1 illustrates the data extraction windows for cases and controls.

Figure 5.1 EHR cohort identification and candidate covariates selection




### 5.2.3 Methodology Framework

Figure 5.2 Methodological Framework




As illustrated in Figure 5.2, our methodology comprises three steps: (1) Obtaining the confounding adjusted signal score for each drug-ADR pair from individual health data; (2) Calibrating the signal scores based on the empirical distribution derived from a set of reference negative controls; (3) Combining calibrated signal scores from disparate databases. In what follows, we elaborate the technical details in each of the three steps.

**5.2.3.1 Obtaining confounding-adjusted ADR signal scores**

For FAERS and NYP/CUMC EHR, we generated signal scores, which are signified by one-sided p-values, using the adjusted log odds ratios (log ORs) and their standard errors calculated by the equations 4.1 - 4.3.

For GE EHR and claims data, the signal scores (one-sided p-values) were generated based on the log relative risks (log RRs) and their standard errors provided by their optimal methods.

#### 5.2.3.2 Standardizing ADR signal scores using p-value calibration

If there is no drug-ADR association, the signal scores using one-sided p-value should be uniformly distributed over the interval (0, 1) in theory. In reality, that is often deviated and leads to an inflated false discovery rate. We apply the estimation algorithm to a set of negative controls in the reference standard, and estimate the empirical distribution of resulting signal scores following formula (5.1), where $q_i$ represents a one-sided p-value of a negative control and n represents the number of negative controls in the reference standard. $\widehat{F_n}(x)$ is then used as the null distribution to calibrate signal scores. This calibration was ADR specific by assuming that signal scores within similar groups have their inherent ranking. For example, a negative control for ALI was not considered in the calibration of AMI. This procedure could be considered as a supervised training procedure with the training set consisted of negative controls in the reference standard. Since we did not use the overall reference standard for both training data and testing data, over-fitting is less of a problem.

Equation 5.1 P-value adjustment using empirical distribution based on negative controls of reference standard

$$\widehat{F_n}(x) = \frac{1}{n}\sum_{i=1}^{n} I\{x < q_i\}$$

#### 5.2.3.3 Combining ADR signals from two heterogeneous databases.

Let $p_{i1}$ denote the i$^{th}$ ADR signal-score computed from source 1, (e.g. the NYP/CUMC EHR), and $p_{i2}$ denote the signal-score for the same drug-ADR pair computed from source 2 (e.g. FAERS). We used the formula 5.2 to combine the signal scores from the two data sources.

Equation 5.2 The composite signal calculated based on two equally contributed signals

$$-2 * [\log(p_{i1}) + \log(p_{i2})] \sim \chi^2_{(4)} \; under\ the\ null\ hypothesis$$

#### 5.2.3.4 Generalizing the combined method

In this work, we combined the signal scores from multiple data sources with equal weights. This approach could be generalized to weighted combination. Formula 5.3 was used to compute a weighted combined signal, where the weights are proportional to their precision associated with the data set so that more weight was assigned when signal scores were more precise.

Equation 5.3 The composite signal calculated based on two weightily contributed signals

$$-2 * [\varpi_i \log(p_{i1}) + (1 - \varpi_i) * \log(p_{i2})] \sim \chi^2_{(df*)} \; under\ the\ null\ hypothesis$$

$$with\ \varpi_i = \left(Var\left(\beta_{i1}^{(2)}\right)\right)^{-1} \Big/ \left(\left(Var\left(\beta_{i1}^{(2)}\right)\right)^{-1} + \left(Var\left(\beta_{i2}^{(2)}\right)\right)^{-1}\right)$$

### 5.2.4 Evaluation Design

We used the reference standard developed by OMOP as described above to generate three reference standards for our study. For reference standard 1, we restricted the evaluation to those drug ADR pairs for which FAERS contained at least one case report and the NYP/CUMC EHR contained at least five patients who were exposed to the studied medications and who were later diagnosed with the studied ADR. For reference standard 2, we restricted the evaluation to those drug-ADR pairs for which FAERS had at least one case report and the GE EHR had results available in the OMOP result set. For reference standard 3, we restricted the evaluation to those

drug ADR pairs for which FAERS had at least one case report and the CCAE had results available in the OMOP result set.

Based on reference set 1, 2 or 3, the performance of the combined system was compared against the performance of signal scores generated by each data source independently. Performance was measured using the area under the receiver operator characteristics (ROC) curve (AUC). To test if the differences of AUCs based on the different combination systems were statistically significant, we computed a one-sided p-value for the hypothesis that the difference between the AUC of the two systems was not equal to 0. The tests were computed using a bootstrapping method. To ensure the p-values were computed based on large enough samples of signal-scores, and to get a single answer representing all outcomes, the significant tests were based on overall reference sets used in each experiment.

We further studied the nature and proper use of the combined system on the basis of four scenarios that could occur in actual pharmacovigilance practice where clinical assessors deal with frequently in their routine work. Using the cutoff p-value of 0.05, we considered a drug-ADR pair as a signal if its p-value is less than 0.05. Accordingly, four scenarios are: (1) a drug-ADR pair has p-value $< 0.05$ in both FAERS and healthcare databases meaning a consistent signal is exhibited in both sources, (2) a drug-ADR pair has p-value $\geq 0.05$ in both data sources meaning the lack of this signal in either source, (3) a drug-ADR signal appears in FAERS but not in healthcare database meaning an inconsistent signal is exhibited and (4) a drug-ADR signal appears in healthcare database but not in FAERS also meaning an inconsistent signal is exhibited.

We also compared the AUC before and after confounding adjustment on the basis of the FAERS and NYP/CUMC EHR respectively. Furthermore, we identified false positive signals in

NYP/CUMC EHR by selecting those negative controls that produced a one-sided p-value < 0.05 in the confounding adjustment analysis. We identified false negative signals in EHR by selecting those positive controls that had a one-sided p-value > 0.05 in the confounding adjustment analysis. In addition, we compared the AUC performance of the confounding adjustment method with the cutting-edge method Gamma Poisson Shrinkage (GPS) that produces signal scores signified by lower 5th percentile of the posterior observed-to-expected distribution (EB05) on the basis of FAERS data. The evaluation was restricted to those drug-ADR pairs for which FAERS had at least one case report. Furthermore we assigned a signal score value of 0, lowest possible signal score for EB05, to those drug-ADR pairs that were never mentioned as primarily suspected relationships, which consequently were not included in the analysis using GPS.

**5.3 Results**

We used 2.7 million case reports from FAERS, 0.3 million patients from the NYP/CUMC EHR, 11 million patients from the GE EHR data and 47 million patients from the CCAE claims data. Some case reports were excluded from FAERS due to typos of drug names or/and the incomplete list of drug names using STITCH.

Table 5.1 listed the number of positive and negative controls for the four ADRs when combining FAERS and the NYP/CUMC EHR, FAERS and the GE EHR, and FAERS and claims data respectively.

Table 5.2 shows the AUCs with and without confounding adjustment, which suggests that the confounding adjustment was essential for both FAERS and NYP/CUMC EHR individually. Moreover, the AUCs after the confounding adjustment in FAERS were statistically significantly

better than those without the adjustment. However, we did not observe a substantial improvement when using the NYP/CUMC EHR. In total, there were 4 false positive signals and 35 false negative signals for the NYP/CUMC EHR. We display them correspondently in Table 5.3 and Table 5.4.

Table 5.1 Subsets of the OMOP reference standard used in the three experiments

| | Reference Set 1 FAERS & NYP/CUMC EHR | | Reference Set 2 FAERS & GE EHR | | Reference Set 3 FAERS & Claims data | |
|---|---|---|---|---|---|---|
| | P | N | P | N | P | N |
| Acute Renal Failure | 16 | 37 | 21 | 48 | 21 | 51 |
| Acute Liver Injury | 52 | 16 | 75 | 30 | 77 | 32 |
| Acute Myocardial Infarction | 10 | 28 | 33 | 51 | 33 | 58 |
| Upper GI Bleed | 17 | 38 | 24 | 57 | 24 | 63 |
| Total | 95 | 119 | 153 | 186 | 155 | 204 |

P: positive controls in the reference standard; N: negative controls

Table 5.2 AUC for FAERS and NYP/CUMC EHR before and after confounding adjustment

| ADR | FAERS | | NYP/CUMC EHR | |
|---|---|---|---|---|
| | Unadjusted | Adjusted | Unadjusted | Adjusted |
| Acute Renal Failure | 0.50 | 0.89 | 0.58 | 0.61 |
| Acute Liver Injury | 0.50 | 0.70 | 0.55 | 0.45 |
| Acute Myocardial Infarction | 0.48 | 0.65 | 0.44 | 0.53 |
| Upper GI Bleed | 0.49 | 0.83 | 0.48 | 0.54 |
| Total | 0.49 | 0.75 | 0.55 | 0.51 |

Unadjusted: signal scores (one-sided p-values) are not adjusted for the confounding effect
Adjusted: signal scores (one-sided p-values) are adjusted for the confounding effect

Table 5.3 False positive signals in the EHR

| Medication | ADR | a | B | c | D | Pvalue1 | Pvalue2 |
|---|---|---|---|---|---|---|---|
| hyoscyamine | GI Bleed | 24 | 976 | 27614 | 2094848 | 0.00 | 0.00 |
| rosiglitazone | GI Bleed | 213 | 30123 | 27425 | 2065701 | 0.00 | 0.01 |
| hyoscyamine | ALI | 19 | 981 | 39029 | 2083433 | 0.00 | 0.01 |
| metaxalone | AMI | 81 | 2214 | 18371 | 2102796 | 0.09 | 0.04 |

a: number of patients exposed to the medication who developed the ADR
b: number of patients exposed to the medication who did not develop the ADR
c: number of patients not exposed to the medication who developed the ADR
d: number of patients not exposed to the medication who did not develop the ADR
pvalue1: unadjusted one-sided p-value
pvalue2: adjusted one-sided p-value

Table 5.4 False negative signals in the NYP/CUMC EHR

| Medication | ADR | a | B | c | d | Pvalue1 | Pvalue2 |
|---|---|---|---|---|---|---|---|
| amlodipine | AMI | 963 | 17010 | 4854 | 191352 | 0.00 | 1.00 |
| darbepoetin alfa | AMI | 134 | 2160 | 5683 | 206202 | 0.00 | 1.00 |
| dipyridamole | AMI | 102 | 1510 | 5715 | 206852 | 0.00 | 0.99 |
| nifedipine | AMI | 209 | 3490 | 5608 | 204872 | 0.00 | 0.17 |
| Acyclovir | ARF | 266 | 2631 | 14624 | 197082 | 0.00 | 1.00 |
| allopurinol | ARF | 725 | 2079 | 14165 | 197634 | 0.00 | 0.19 |
| Captopril | ARF | 400 | 1739 | 14490 | 197974 | 0.00 | 1.00 |
| cyclosporine | ARF | 352 | 907 | 14538 | 198806 | 0.00 | 1.00 |
| enalaprilat | ARF | 228 | 1191 | 14662 | 198522 | 0.00 | 0.85 |
| Ibuprofen | ARF | 756 | 32402 | 14134 | 167311 | 1.00 | 1.00 |
| Ketorolac | ARF | 164 | 5386 | 14726 | 194327 | 1.00 | 1.00 |
| Lisinopril | ARF | 2815 | 16984 | 12075 | 182729 | 0.00 | 0.98 |
| meloxicam | ARF | 103 | 1977 | 14787 | 197736 | 1.00 | 0.90 |
| Naproxen | ARF | 256 | 6767 | 14634 | 192946 | 1.00 | 1.00 |
| allopurinol | ALI | 164 | 2926 | 5935 | 203323 | 0.00 | 1.00 |
| ciprofloxacin | ALI | 222 | 4892 | 5877 | 201357 | 0.00 | 1.00 |
| cyclosporine | ALI | 178 | 1117 | 5921 | 205132 | 0.00 | 0.94 |
| Diltiazem | ALI | 224 | 5814 | 5875 | 200435 | 0.00 | 1.00 |
| fluconazole | ALI | 330 | 4845 | 5769 | 201404 | 0.00 | 1.00 |
| Ibuprofen | ALI | 545 | 30766 | 5554 | 175483 | 1.00 | 1.00 |
| Ketorolac | ALI | 125 | 5120 | 5974 | 201129 | 0.98 | 1.00 |
| lamivudine | ALI | 126 | 1204 | 5973 | 205045 | 0.00 | 1.00 |
| levofloxacin | ALI | 591 | 11486 | 5508 | 194763 | 0.00 | 1.00 |
| Lisinopril | ALI | 738 | 19117 | 5361 | 187132 | 0.00 | 1.00 |
| Naproxen | ALI | 134 | 5921 | 5965 | 200328 | 1.00 | 1.00 |
| nifedipine | ALI | 150 | 3742 | 5949 | 202507 | 0.00 | 0.98 |
| Ramipril | ALI | 139 | 3562 | 5960 | 202687 | 0.00 | 0.99 |
| citalopram | GI BLEED | 246 | 4250 | 6437 | 202220 | 0.00 | 1.00 |
| clopidogrel | GI BLEED | 542 | 12940 | 6141 | 193530 | 0.00 | 1.00 |
| escitalopram | GI BLEED | 188 | 3616 | 6495 | 202854 | 0.00 | 0.72 |
| Ibuprofen | GI BLEED | 492 | 27177 | 6191 | 179293 | 1.00 | 1.00 |
| Ketorolac | GI BLEED | 105 | 4813 | 6578 | 201657 | 1.00 | 1.00 |
| Naproxen | GI BLEED | 168 | 5052 | 6515 | 201418 | 0.36 | 1.00 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| potassium chloride | GI BLEED | 154 | 2778 | 6529 | 203692 | 0.00 | 1.00 |
| Sertraline | GI BLEED | 256 | 4850 | 6427 | 201620 | 0.00 | 0.85 |

a: number of patients exposed to the medication who developed the ADR
b: number of patients exposed to the medication who did not develop the ADR
c: number of patients not exposed to the medication who developed the ADR
d: number of patients not exposed to the medication who did not develop the ADR
pvalue1: unadjusted one-sided p-value
pvalue2: adjusted one-sided p-value

The results from experiment 1 are presented in Table 5.5. We found that the FAERS system performed significantly better than the NYP/CUMC EHR system. Combining FAERS and NYP/CUMC EHR data did not improve the ADR detection performance of FAERS, although it did not harm it either. The combined system still performed significantly better than the NYP/CUMC alone. Experiment 2, which is also presented in Table 5.5, shows that the combined system outperformed both the FAERS and the GE EHR individual systems. Improvements were observed for all the outcomes, although at different levels. The AUC of the combined system ranged from 76% for ALI to 92% for ARF. For individual systems, the GE EHR system had better AUC performance for AMI than FAERS, but worse than FAERS for ARF, ALI and GIB. Similar results were found when combining FAERS with the CCAE in experiment 3. The CCAE had better performance than FAERS for AMI and GIB, but was worse for the other two. Again, the combined system outperformed the individual ones for all the four outcomes.

Table 5.5 AUC of signal detection performance for FAERS, healthcare data and combined systems

| | Experiment 1. Combining FAERS and NYP/CUMC EHR | | |
|---|---|---|---|
| ADR | FAERS | EHR | Combined |
| Acute renal failure | 0.89 | 0.61 | 0.89 |
| Acute liver injury | 0.70 | 0.45 | 0.68 |
| Acute myocardial infarction | 0.65 | 0.53 | 0.70 |
| Upper GI bleeding | 0.83 | 0.54 | 0.83 |
| Total | 0.75 | 0.51 | 0.74 |
| | | | |
| | Experiment 2. Combining FAERS and GE EHR | | |
| ADR | FAERS | GE | Combined |
| Acute renal failure | 0.91 | 0.68 | 0.92 |
| Acute liver injury | 0.71 | 0.63 | 0.76 |
| Acute myocardial infarction | 0.72 | 0.80 | 0.82 |
| Upper GI bleeding | 0.80 | 0.77 | 0.87 |
| Total | 0.76 | 0.76 | 0.82 |
| | | | |
| | Experiment 3. Combining FAERS and the claims data | | |
| ADR | FAERS | Claims | Combined |
| Acute renal failure | 0.91 | 0.83 | 0.93 |
| Acute liver injury | 0.72 | 0.69 | 0.79 |
| Acute myocardial infarction | 0.71 | 0.77 | 0.82 |
| Upper GI bleeding | 0.81 | 0.83 | 0.86 |
| Total | 0.76 | 0.78 | 0.82 |

Results in Table 5.6 shows that the combined system achieved better AUC performances in most of four scenarios for two of the combination studies. Specifically, when compared with the better performing individual system, the combined system increased AUC improvement ranging from 3% to 11% although it decreased the one of 7% whereas the signals exhibit in FAERS but not in claims database. The difference of AUC performance was defined as the AUC of combined system minus the AUC of better performing individual system.

Table 5.6 The AUC performance of FAERS, healthcare data and the combined system on the basis of four scenarios

| | Consistent information in two sources | | Inconsistent information in two sources | |
|---|---|---|---|---|
| Scenarios | Both FAERS and healthcare database show signals | Neither FAERS nor healthcare database show signals | FAERS shows signal but healthcare database does not | Healthcare database shows signal but FAERS does not |
| Positive/negative controls* | 25/0 | 61/152 | 29/11 | 38/23 |
| FAERS alone | NA | 0.71 | 0.73 | 0.60 |
| GE alone | NA | 0.69 | 0.78 | 0.68 |
| FAERS and GE combined | NA | 0.75 | 0.89 | 0.68 |
| | | | | |
| Positive/negative controls* | 49/3 | 16/104 | 7/8 | 83/89 |
| FAERS alone | 0.84 | 0.68 | 0.68 | 0.67 |
| Claims alone | 0.69 | 0.50 | 0.86 | 0.67 |
| FAERS and Claims combined | 0.89 | 0.74 | 0.79 | 0.68 |

*Positive and negative controls are defined according to the reference standard
Signals are identified based on one-sided p-value < 0.05
NA: AUC performances are not computable when only positive controls are available

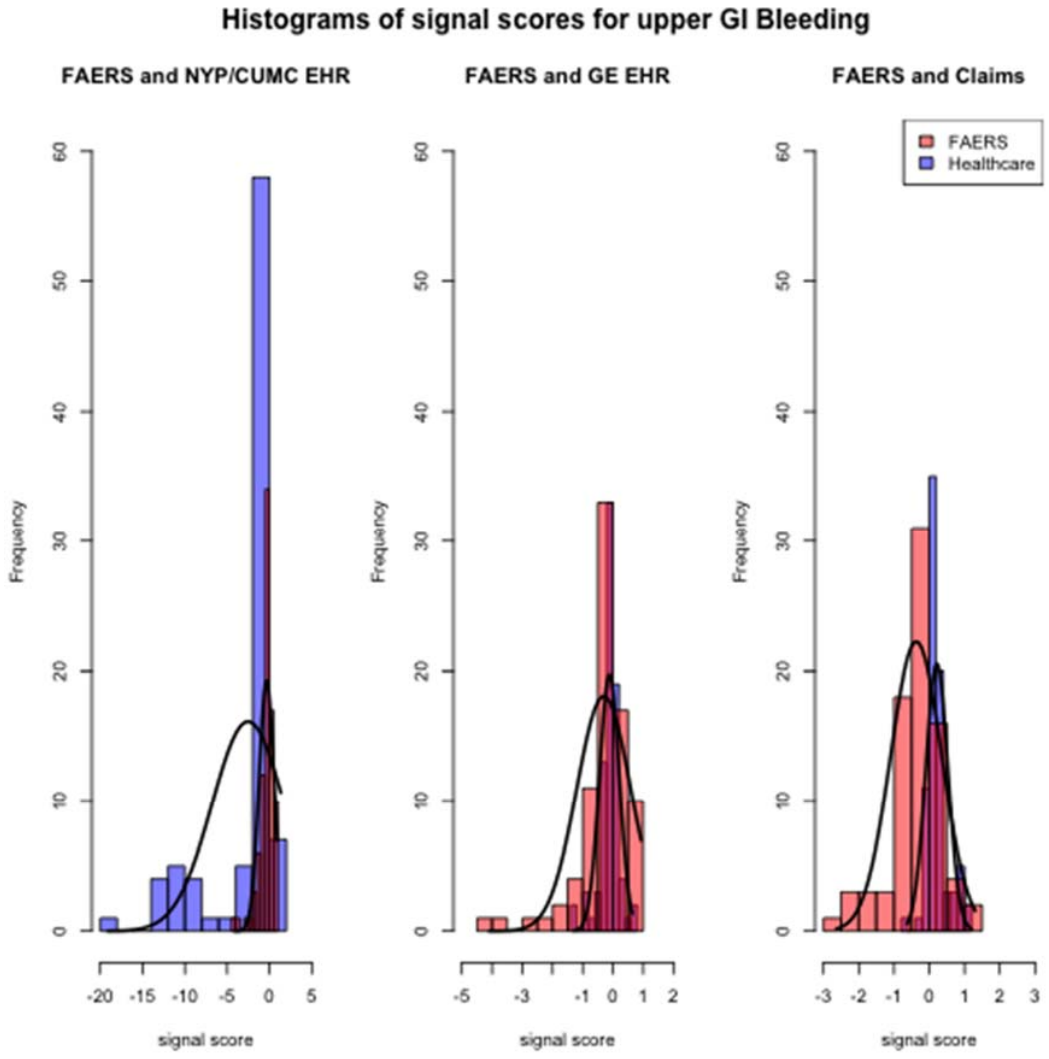
Using the cutoff p-value of 0.05, we evaluated the precision and recall of the two combined systems – the combination system using FAERS and GE EHR, and the combination system using FAERS and claims data. Combining FAERS with the GE EHR resulted in higher recall (0.41 versus 0.35), while the precisions of the two combination systems were almost identical (0.925 versus 0.931). Using the same cutoff p-value, eight more signals were detected only by the combined system, as shown in Table 5.7. Among them, seven of the eight were true positive signals.

Table 5.7 ADR signals detected only using the combined GE and FAERS system and their one-sided p-values in three systems

| medication | ADR | Ground Truth | FAERS | GE | Combined system |
|---|---|---|---|---|---|
| piroxicam | ARF | 1 | 0.299 | 0.432 | 0.043 |
| amoxapine | AMI | 1 | 0.076 | 0.118 | 0.007 |
| diflunisal | AMI | 1 | 0.109 | 0.192 | 0.007 |
| eletriptan | AMI | 1 | 0.682 | 0.072 | 0.034 |
| nabumetone | AMI | 1 | 0.079 | 0.494 | 0.035 |
| nelfinavir | AMI | 0 | 0.292 | 0.263 | 0.044 |
| zolmitriptan | AMI | 1 | 0.224 | 0.381 | 0.034 |
| ketorolac | GIB | 1 | 0.425 | 0.069 | 0.041 |

Figure 5.3 shows the histograms of the signal scores for upper GI bleeding in each experiment. It is apparent from the figure that the scale of signal score for FAERS did not overlap substantially with each healthcare data set, and the distribution of the signal scores did not follow a normal distribution.

Figure 5.3 Histograms of signal scores when combining FAERS with the three healthcare data sets




Signal scores for FAERS and the EHR are signified by log OR, and signal scores for the GE EHR and the claims data are signified by log RR.

## 5.4 Discussion

The main results of our evaluation show that combining signals from two relatively large data sources (e.g. FAERS and the GE EHR data, FAERS and the claims data) using the proposed methodological framework led to an overall significant improvement, which was replicated for the different outcomes. However, we did not observe the improvement when combining FAERS with the NYP/CUMC EHR. The discrepancies are possibly attributed to issues such as small data size and sample biases.

### 5.4.1 Small data size

NYP/CUMC EHRs have already been successfully used for detecting safety signals in several studies (Schneeweiss, Rassen et al. 2009, Coloma, Schuemie et al. 2011, Gagne, Glynn et al. 2012). However, challenges remain because of the relatively small size of the data. There were only 0.3 million patients in NYP/CUMC EHR compared with 11 million in GE EHR and 47 million in the claims data. Since ADRs generally occur infrequently in the EHRs, and their signals are often weak, a large data size is essential for effective detection.

For the same reason, we could clearly observe that higher prevalence of an ADR resulted in better performance on the basis of the NYP/CUMC EHR. Specifically, NYP/CUMC EHR included 14,890 patients having ARF, 6,099 patients having ALI, 5,817 patients having AMI, and 6,683 patients having upper GI Bleed. ARF with many more patients, almost three times as many patients than those developing AMI, had better AUC performance than the other three ADRs. Furthermore, when using NYP/CUMC EHR to detect the drugs associated with ARF, we achieved 100% precision, and successfully identified three true positive medications, hydrochlorothiazide, telmisartan and candesartan.

### 5.4.2 Sampling biases

The NYP/CUMC EHR data came from a tertiary care academic medical center in a major metropolitan inner city area, which may lead to a highly skewed population. In addition, many of the patients included in our analysis could have been referred from other facilities and therefore their EHR data may have been incomplete because it may have lacked longitudinal information for many of those patients. Namely, only 37% of patients had at least one outpatient visit and only 14% of patients had more than one visit. Moreover, NYP/CUMC EHR data was not linked to pharmacy prescriptions or refills, and the medications extracted from free-text notes were just the mentions of medications, and therefore temporal relationships between medication exposures and ADR events may not have been definitive. In contrast, the GE EHR represented a large outpatient population and captured longitudinal patient information, such as ICD-9 coded medical problems and prescriptions. The claims data represented a much larger and more diverse population, and captured longitudinal patient information including diagnosis codes for billing purposes, as well as dates when prescriptions were filled or refilled. However, both the GE EHR and the claims data may also have faced the challenge of a skewed patient population, such as sicker patients having much more visits, and more prescriptions and refills in the database (Ryan, Madigan et al. 2012).

### 5.4.3 Usefulness for pharmacovigilance practice

The AUC evaluation showed that FAERS had substantially better performance for ARF and ALI, and worse performance for AMI than healthcare data, which indicates that no single source may provide best evidence for all ADR detections. Therefore, synthesis of evidence from multiple streams of information is extremely significant. Currently, clinical assessors carry out

the analysis of evidence from multiple sources. For example, clinical assessors may validate or want to evaluate a signal from different resources, as those generated from SRSs and/or healthcare data. Thus, a common situation that clinical assessors need to deal with is inconsistent or conflicting information from the different data sources.

Results show that the combined ADR signals generated by the proposed method improved the AUC performance significantly compared with individual systems. In addition, we evaluated the combined system in four scenarios: (1) signals were demonstrated in both FAERS and the healthcare data; (2) no signals were generated in FAERS and in the healthcare data; (3) signals only appeared in FAERS but not in healthcare database and (4) signals only appeared in healthcare database but not in FAERS. We observed the consistent improvement was achieved by the combined system except for when signals appeared in FAERS but not in claims database. However, the combined system was still better than FAERS system alone in ranking potential signals. Therefore, the proposed system could serve as a tool for clinical assessors when they review ADR cases. For example, in the scenario of consistent signals, clinical assessors are more likely to believe the existence of the signals and may want to select the strongest signals for further assessment; the combined system could prioritize signals by integrating the two sources. In the scenario of inconsistent signals, the combined system is able to resolve inconsistent or conflicting statistical information and then provide a single response through the consolidation of statistical information from the two sources. In the scenario where no single source provides a signal, the combined system could possibly transfer two relatively weak signals into a stronger composite one. For instance, eight more signals were detected only using the combined GE and FAERS system and seven of them were true positives, which is promising. However, a practical challenge is how to effectively communicate these results to the clinical assessors. In addition,

combining FAERS with GE resulted in higher recall and almost identical precision when compared with combining FAERS with claims data. We also observed that healthcare databases were more sensitive for ADR detection than FAERS in terms of identifying more signals. We acknowledge that the recall and precision are threshold-dependent performance metrics. Hence the results may vary when using different thresholds.

### 5.4.4 Related work

Our method was designed originally to combine the NYP/CUMC EHR with FAERS, which is the first such study. Harpaz et al designed an empirical Bayes model to combine signals across FAERS and claims data showing its effectiveness using the same reference standard used in this study. However, that method required that the data satisfy two assumptions: a) the signal-scores generated from each individual data source should be on approximately the same scale, and b) the scores should follow the log normal distribution. Our data sets did not meet these assumptions. Figure 5.3 illustrates the violation of the above two assumptions for upper GI bleeding, but the other three ADRs had similar results.

### 5.4.5 Methods to deal with confounding

The capability to reduce or eliminate confounding is a major aim of ADR detection. Self-controlled designs have recently been proposed and successfully utilized in ADR detection based on longitudinal healthcare data. They attempt to identify equivalent periods of unexposed time within the same patients, against which to compare the same patients' exposed time. However, NYP/CUMC EHR data lacked this kind of longitudinal information relating to when a patient was put on or taken off a medication. Our prior study showed that insufficient confounder

selection led to high false positive rates (Li, Salmasian et al. 2013) and therefore we designed the two-step LASSO regression (step 1 of the proposed methodological framework) to select more associated confounders. The AUC performances were generally improved after the confounding adjustment except for ALI. We also applied this algorithm to the FAERS data and the AUC performances were statistically significantly better with this algorithm than without it. Furthermore, results in Table 5.8 shows that the confounding adjustment method achieved comparable performance with the cutting-edge algorithm GPS based on FAERS. For example, the confounding adjustment method had better AUC performances in ARF, AMI and GIB, and lower AUC performance in ALI. The advantage of the two-step LASSO compared with the single LASSO is shown in Figure 5.4 and Figure 5.5, where the two-step LASSO separated positive controls more from negative ones, reduced the false positive rate and achieved better AUC performance.

Table 5.8 Reference set and the AUC performance for the confounding adjustment method and Gamma Poisson Shrinkage (GPS) method on the basis of FAERS from 2004 to 2010

| | Reference set | | AUC performance | |
|---|---|---|---|---|
| | Positive | Negative | Confounding adjustment method (p-value) | GPS (EB05) |
| Acute renal failure | 23 | 52 | 0.90 | 0.76 |
| Acute liver injury | 77 | 33 | 0.72 | 0.87 |
| Acute myocardial infarction | 38 | 59 | 0.72 | 0.70 |
| Upper GI bleeding | 24 | 63 | 0.81 | 0.79 |

Figure 5.4 Comparison between single LASSO and double LASSO on the basis of the EHR for acute renal failure

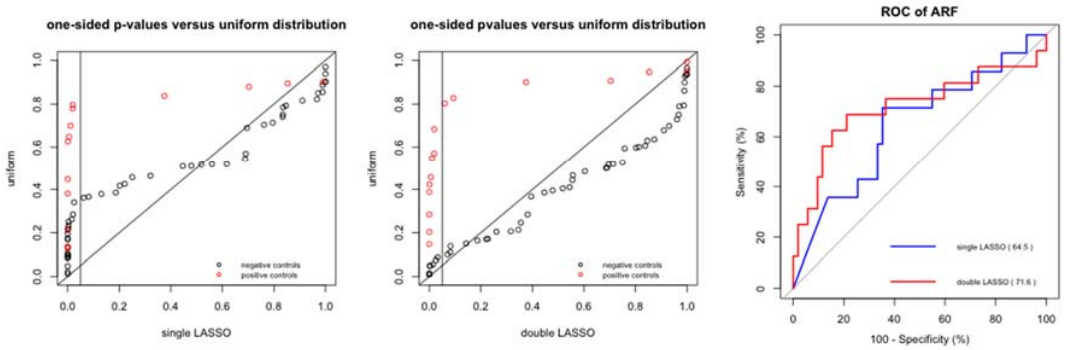
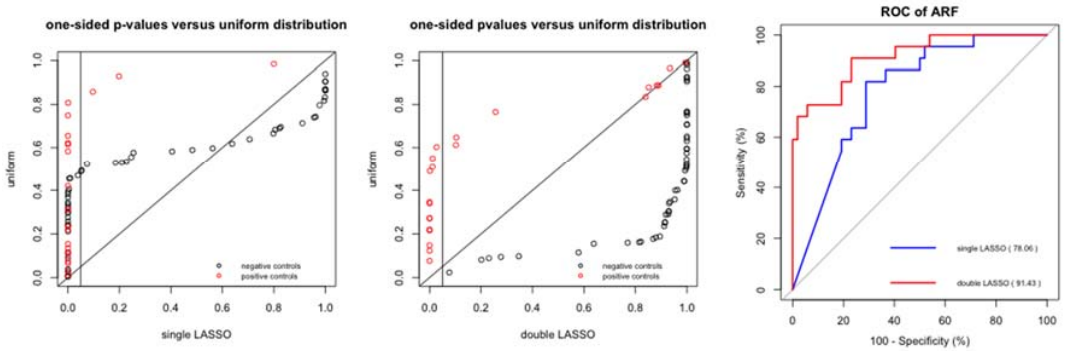



Figure 5.5 Comparison between single LASSO and double LASSO on the basis of the FAERS for acute renal failure




### 5.4.6 False positive signals in CUMC/NYP EHR

One false positive signal was rosiglitazone for GI Bleed. Rosiglitazone was mentioned on the records of 1,587 patients where 133 of the patients developed GI Bleed. The confounding adjustment method scored this pair with a one-sided p-value of 0.01. In contrast, pioglitazone is in the same drug class, and was mentioned in the records of 2,477 patients where 110 patients developed GI bleed. The confounding adjustment method scored this pair with a one-sided p-value of 1.

### 5.4.7 False negative signals in CUMC/NYP EHR

Ryan et al demonstrated that the cohort method using high dimensional features selected by Bayesian logistic regression generally yielded a negatively biased estimate (Ryan, Stang et al. 2013). We observed the same trend in our data set and summarize possible reasons for false negative signals: 1) Data sparseness since there were not enough patients exposed to the studied medications while the ADR occurrences were quite rare. 2) Confounding by indication. It is common that the indication for a drug may bias the estimated association if it is associated with an increased risk of the ADR itself. For example, amlodipine and nifedipine have hypertension as an indication, but hypertension was also related to AMI, and therefore did not produce positive associations. However, amlodipine and nifedipine were in the reference standard as being positive for AMI. The proposed method could not deal with this issue correctly, and more clinical knowledge may be needed. For example, we may compare a medication with the other medications having the same treatment regime to better understand its relationship with the ADR.

### 5.4.8 Generalized the combined method

The results of Table 5.9 demonstrated that equally combined strategy outperformed the weighted combined strategy. However, the performance may vary when apply to different signal scores generated from different data sources. The value of using equally combined strategy is that it overcomes the problem existing in the weighted combined strategy when one of the data sources being considered is much larger than the others, in which case it may dominate the weighting for certain associations. In addition, the substantial improvement of the combined system after p-value calibration demonstrated that the empirical calibration is needed to correct p-values. The

overall AUC performance for CCAE or GE was worse after the p-value calibration because quite a few calibrated p-values were tied to each other.

Table 5.9 AUC of signal detection performance for the equally and weighted combined systems

| ADR | Unstandardized ADR signal scores | | | | Standardized ADR signal scores | | | |
|---|---|---|---|---|---|---|---|---|
| | Experiment 1. Combining FAERS and the EHR | | | | | | | |
| | FAERS | EHR | Equally | Weighted | FAERS | EHR | Equally | Weighted |
| Acute renal failure | 0.89 | 0.61 | 0.89 | 0.88 | 0.89 | 0.62 | 0.89 | 0.89 |
| Acute liver injury | 0.70 | 0.45 | 0.67 | 0.68 | 0.71 | 0.47 | 0.68 | 0.68 |
| Acute myocardial infarction | 0.65 | 0.53 | 0.70 | 0.67 | 0.66 | 0.53 | 0.70 | 0.68 |
| Upper GI bleeding | 0.83 | 0.54 | 0.84 | 0.82 | 0.84 | 0.54 | 0.83 | 0.81 |
| Total | 0.75 | 0.51 | 0.75 | 0.75 | 0.75 | 0.52 | 0.74 | 0.73 |
| | Experiment 2. Combining FAERS and the GE EHR | | | | | | | |
| | FAERS | GE | Equally | Weighted | FAERS | GE | Equally | Weighted |
| Acute renal failure | 0.91 | 0.68 | 0.87 | 0.80 | 0.91 | 0.69 | 0.92 | 0.88 |
| Acute liver injury | 0.71 | 0.63 | 0.70 | 0.63 | 0.72 | 0.64 | 0.76 | 0.75 |
| Acute myocardial infarction | 0.72 | 0.80 | 0.80 | 0.80 | 0.73 | 0.80 | 0.82 | 0.83 |
| Upper GI bleeding | 0.8 | 0.77 | 0.83 | 0.83 | 0.81 | 0.78 | 0.87 | 0.87 |
| Total | 0.76 | 0.76 | 0.82 | 0.77 | 0.77 | 0.71 | 0.82 | 0.82 |
| | Experiment 3. Combining FAERS and the claims data | | | | | | | |
| | FAERS | Claims | Equally | Weighted | FAERS | Claims | Equally | Weighted |
| Acute renal failure | 0.91 | 0.83 | 0.88 | 0.80 | 0.92 | 0.83 | 0.93 | 0.93 |
| Acute liver injury | 0.72 | 0.69 | 0.72 | 0.67 | 0.73 | 0.70 | 0.79 | 0.83 |
| Acute myocardial infarction | 0.71 | 0.77 | 0.77 | 0.74 | 0.72 | 0.77 | 0.82 | 0.76 |
| Upper GI bleeding | 0.81 | 0.83 | 0.83 | 0.82 | 0.81 | 0.84 | 0.86 | 0.85 |
| Total | 0.76 | 0.78 | 0.81 | 0.76 | 0.77 | 0.73 | 0.82 | 0.80 |

Standardized: signal scores are standardized using p-value calibration with the empirical null distribution derived from negative controls of the reference standard
Equally: equally combined system
Weighted: weighted combined system

### 5.4.9 Limitations

This study had several limitations. First, using the NYP/CUMC was a limitation because of its relatively small population, which limited EHR signal detection capability, and therefore performance of the combined system as well. In future work, we plan to include additional EHR data from multiple sites. Second, the NYP/CUMC data is mainly dominated by inpatient data so that acquiring comprehensive and longitudinal patient information is a challenge. Second, when using the NYP/CUMC EHR, we simply adopted the OMOP outcome definitions, which may not be optimal for the EHR data set, and could have lead to outcome misclassification including both false positive and false negative patients. Third, the confounding adjustment method did not deal well with drugs given only to a particular patient population and therefore the control groups on the basis of a general population were not representative for that population. Therefore selecting patients having the same indications may be more appropriate. Fourth, when using FAERS, we did not remove duplicate reports or correct terminological errors. Lastly, the reference standard consists of test cases that were publicly known during the time frame of our evaluation, and thus the performance may be different when applied to actual clinical care. For example, physicians may be less likely to prescribe those drugs during the study timeframe because they already were aware of the ADRs.

## 5.5 Conclusion

In this paper, we described a method for ADR detection that combined FAERS with healthcare data and showed significant improvement when individual healthcare resources had sufficient amounts of data. Although the small NYP/CUMC EHR database did not contribute to

improvement, use of the large-size network-based GE EHR data and claims data did significantly show improved performance when combined with the FAERs data. An advantage of this method is that it can serve as a tool for synthesizing evidence for clinical assessors in actual pharmacovigilance practice.

# CHAPTER 6 Summary and Conclusions

## 6.1 Summary

This dissertation has proposed and investigated three sequential hypotheses in order to demonstrate that a framework that computationally integrates ADR evidence from multiple data sources is feasible and efficient for ADR detection. The three hypotheses are: (1) a data-driven method could achieve comparable or slightly higher accuracy than those trained with experts and existing automatic methods; (2) a data-driven and regression-based method using primary suspected, secondary suspected, concomitant medications and indications could accomplish comparable or slightly better accuracy than the cutting edge algorithm Gamma Poisson Shrinkage (GPS) using primary suspected medications only; (3) a computational integration method could result in a more accurate ADR detection system than individual systems based on either FAERS or observational healthcare data.

The first hypothesis involved research concerning a data-driven and regression-based method for alleviating confounding effect in ADR detection using a single facility EHR database. Results showed that precision of ADR detection was 83.3% for rhabdomyolysis and 60.8% for pancreatitis when using the proposed method, and it identified several drug safety signals that are interesting for further clinical review. Compared with four comparators, the proposed method achieved either a higher or similar precision, and had the unique ability to provide insight into confounders for each specific medication–ADR pair, and could be easily adapted to other EHR systems.

The second hypothesis improved the performance of ADR detection by using information of concomitant medications and indications in additional to primary suspected medications commonly utilized by the disproportionality analysis. We proposed the use of two-step Lasso regression to select sufficient confounders and then to estimate the associations when adjusting for these confounders. The overall goal was to alleviate the confounding effect due to co-medications and indications. Evaluation indicated that two-step Lasso regression improved AUC for acute renal failure and acute myocardial infarction, but not for the other two, compared with the well-established Gamma Poisson shrinkage method. Since the reference standard is explicitly correlated with GPS for negative controls, the performance of GPS is over estimated.

Finally, the third hypothesis involved a computational method of integrating ADR evidence from multiple data sources, which was explored and resulted in a more accurate ADR detection system. In the study, we proposed an algorithm to combine signal scores of one-sided p-value mined from SRS and observational healthcare databases. Evaluation based on reference standards consisting of known positive and negative test cases indicated that although there was no improvement in the AUC when combining the SRS and small-scale EHR, the AUC of the combined SRS and large-scale EHR was 0.82 whereas it was 0.76 for each of the individual systems. Similarly, the AUC of the combined SRS and claims system was 0.82 whereas it was 0.76 and 0.78 respectively for the individual systems. Furthermore, the combined system can serve as a tool in actual pharmacovigilance practice to either consolidate consistent information or resolving inconsistent information provided by different data sources, and the combined system is able to detect more signals than any individual system.

## 6.2 Contributions

### 6.2.1 Contributions to Biomedical Informatics

1. The major contribution of this dissertation research to biomedical informatics is that it develops a framework to combine ADR evidence from multiple data sources and demonstrates the framework significantly improved performance of ADR detection using three experiments involving combining the SRS with a single facility EHR, a larger-scale network-based EHR, and a much larger-scale healthcare claims database.

2. The proposed framework requires obtaining and measuring signal scores using the most appropriate methods for each individual data sources and are in the form of one-sided p-value. We developed a data-driven and regression-based method to acquire signal scores from both a single facility EHR system and FAERS, which are proven to be better or comparable with other existing methods.

3. The method to combine ADR data from multiple data sources can be generalized to integrate evidence from other data sources. For example, using biomedical literature and disproportionality analysis to generate ADR signals is another promising direction. The signal scores could be transferred into one-sided p-values and be integrated into the combinational method.

4. The proposed LASSO and two-step LASSO methods together with the feature screening are data-driven methods so that they could be easily adapted to other EHR systems for ADR detection. Moreover, the methods can be generalized to other tasks such as discovering drug off-label use on the basis of EHR data.

### 6.2.2 Contributions to Healthcare

The primary contribution of this dissertation research to healthcare is that it develops an automated system for computationally integrating ADR evidence from heterogeneous sources improving performance of ADR detection. Furthermore this combined system can serve as a tool for assisting manually review process. The synthesis of evidence from multiple streams of information has been an integral part of pharmacovigilance. Yet, it is currently carried out by clinical assessors on an ad hoc basis, in a rather qualitative manner, and usually after a signal is generated. The proposed system can prioritize ADR signals when signals are demonstrated consistently in multiple data sources, resolve conflict information when signals are demonstrated in one data source but not in the other data source, and detect ADR signals that are not able to be detected using individual data sources.

### 6.3 Innovations

The proposed approach is the first real attempt to computationally integrate ADR evidence from SRS and EHR databases in the field of pharmacovigilance.

1. To the best of my knowledge, this is the first time that a computational method of integrating ADR evidence, from an SRS and an EHR, has been developed for achieving higher accuracy of ADR detection compared with individual systems in pharmacovigilance.
2. For the first time, the evaluation has been conducted on the basis of scenarios where two data sources provide consistent or inconsistent information.
3. It is novel that data-driven and regression-based methods have been developed to identify confounders and estimate drug-ADR associations when adjusting for these confounders.

### 6.4 Limitations

Some specific limitations have been described in previous chapters, and are summarized below:

1. The NYP/CUMC EHR data is one of major data sources used in this dissertation, however, the performance of ADR detection when combining the FAERS with this particular EHR system was not improved. The reason is that the NYP/CUMC EHR data came from a tertiary care academic medical center in a major metropolitan inner city area, which may lead to a highly skewed population. In addition, many of the patients included in our analysis could have been referred from other facilities and their EHR data may have been incomplete because it may have lacked longitudinal information for many of those patients. Therefore, temporal information, which is critical for drug-ADR detection, is not well represented in the collected data. Moreover, NYP/CUMC EHR data was not linked to pharmacy prescriptions or refills, and the medications extracted from free-text notes were just the mentions of medications, and therefore temporal relationships between medication exposures and ADR events may not have been definitive. In addition, NYP/CUMC EHR data is relatively small compared with other data sources.
2. The method involving signal score calibration currently relies on the availability of reference negative controls for a particular ADR, which is hard to be generalizable to other ADRs.
3. The development and evaluation of evidence-based methods require benchmarks for signal detection performance, and therefore we used published reference sets to evaluate all proposed systems. However, it should be noted that the reference standards consisted of established ADRs without evidence for or against causal associations with a drug, and the emerging ADRs are different from established ADRs in nature. Therefore, the

performances evaluated based on the reference standards used in this dissertation are retrospective but not prospective which could possibly yield different conclusions.

4. The proposed combined system delivers the evidence in a very concise format, namely, a single index ranging from 0 to 1. However, there is not an established threshold for this index to determine whether a particular drug ADR pair is a potential signal. In addition, although the combined system could discover ADR signals that were missed by individual systems, a real-world challenge is about how to communicate the results with clinical assessors using this purely statistic based model.

### 6.5 Future directions

The research presented in this dissertation provides the initial attempt for developing a framework of integrating ADR evidence from multiple data streams. Future research can extend the current work in various directions.

1. For the data, we know that data size and data quality are extremely important for detecting ADRs, for example, we did not observe the improvement when combining FAERS and an EHR from a single facility. Therefore, we plan to acquire more longitudinal information within a single EHR and collaborate with other facilities for acquiring more EHR data. The common data model developed by OMOP could be helpful for the integration of EHR data.
2. Currently, the method involving signal calibration relies on the availability of reference negative controls. In the future, we plan to develop a theoretical null distribution to calibrate signal scores. In addition, the OMOP has a standard procedure for examining

and collecting negative controls for a particular ADR, which could be generalizable to other ADRs.

3. The confounding adjustment method assumed a single and homogeneous odds ratio for a drug-ADR combination, which may not be appropriate. In future work, we plan to apply clustering algorithms to group patients with similar symptoms or diagnoses and then acquire associations within these relatively homogeneous patient groups. For this study, grouping patients based on their ICD-9 diagnosis codes could be a good start.
4. The confounding method did not deal with drugs given only to a particular patient population and therefore the control groups on the basis of a general population were not representative for that population. In future work, we plan to analyze particular drug-ADR combinations in the subgroup of general population such as elderly patients. In addition, we plan to introduce a knowledgebase involving drug indication information, and compare the studied drug with drugs having the same indication.
5. A more prospective evaluation involving emerging ADR signals will be undertaken in the future using a time-indexed reference standard.
6. The combined system could be extended to incorporate information from more than two data sources, for example, combining signals from the SRS, claims and EHR databases could be next steps. Another promising direction is discovering signals on the basis of a resource that is sensitive for ADR detection, and then re-ranking signals using a resource, which provides high precision.

## 6.6 Conclusions

Completing safety profiles over the market life of a drug is a constant challenge in the field of pharmacovigilance and is critical for patient safety. The research presented in this dissertation has produced several novel findings that provide new insights that demonstrate that computationally integrate ADR evidence from multiple data streams has the potential to detect ADRs earlier. To the best of my knowledge, this is the first study integrating data from a SRS and EHR databases. In conclusion, this dissertation develops a combined system to synthesize evidence and a method to reduce confounding effect in both EHRs and SRSs, which could potentially unveil drug safety profiles and novel adverse events in a timely fashion.

**References**

"FDA Adverse Event Reporting System (FAERS)." from http://www.fda.gov/cder/aers/default.htm.

Master Drug Data Base (MDDB) Documentation Manual. Indianapolis, IN: MediSpan Inc., 1995.

"Medi-Span (R) (Wolters Kluwer Health, Indianapolis, IN) Drug Indications Database (TM) ".

."Medical Dictionary for Regulatory Activities (MedDRA)." from http://www.meddramsso.com.

Micromedex® Health Series [Internet Database]. Greenwood Village, Colo: Thomson Reuters (Healthcare) Inc. Updated periodically. 2010.

"Observational Medical Outcome Partnership (OMOP)." from http://omop.fnih.org.

Ahmad, S. R. (2003). "Adverse drug event monitoring at the Food and Drug Administration." Journal of general internal medicine **18**(1): 57-60.

Alvarez-Requejo, A., A. Carvajal, B. Begaud, Y. Moride, T. Vega and L. M. Arias (1998). "Under-reporting of adverse drug reactions Estimate based on a spontaneous reporting scheme and a sentinel system." European journal of clinical pharmacology **54**(6): 483-488.

Amery, W. K. (1999). "Why there is a need for pharmacovigilance." Pharmacoepidemiology and drug safety **8**(1): 61-64.

Anello, C. and R. T. O'neill (1996). "Does research synthesis have a place in drug regulatory policy? synopsis of issues: assessment of safety and postmarketing surveillance." Clinical Research and Regulatory Affairs **13**(1): 13-21.

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the AMIA Symposium, American Medical Informatics Association.

Ashburn, T. T. and K. B. Thor (2004). "Drug repositioning: identifying and developing new uses for existing drugs." Nature reviews Drug discovery **3**(8): 673-683.

Avillach, P., J.-C. Dufour, G. Diallo, F. Salvo, M. Joubert, F. Thiessard, F. Mougin, G. Trifirò, A. Fourrier-Réglat and A. Pariente (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.

Badalov, N., R. Baradarian, K. Iswara, J. Li, W. Steinberg and S. Tenner (2007). "Drug-induced acute pancreatitis: an evidence-based review." Clinical gastroenterology and hepatology **5**(6): 648-661. e643.

Baeza-Yates, R. and B. Ribeiro-Neto (1999). Modern information retrieval, ACM press New York.

Bate, A., I. Edwards, J. Edwards, E. Swahn, G. Noren and M. Lindquist (2004). Knowledge finding in IMS Disease Analyser Mediplus UK database-effective data mining in longitudinal patient safety data. Drug Safety, ADIS INTERNATIONAL LTD 41 CENTORIAN DR, PRIVATE BAG 65901, MAIRANGI BAY, AUCKLAND 10, NEW ZEALAND.

Bate, A. and S. Evans (2009). "Quantitative signal detection using spontaneous ADR reporting." Pharmacoepidemiology and drug safety **18**(6): 427-436.

Bates, D. W., D. J. Cullen, N. Laird, L. A. Petersen, S. D. Small, D. Servi, G. Laffel, B. J. Sweitzer, B. F. Shea and R. Hallisey (1995). "Incidence of adverse drug events and potential adverse drug events." JAMA: the journal of the American Medical Association **274**(1): 29-34.

Bates, D. W., N. Spell, D. J. Cullen, E. Burdick, N. Laird, L. A. Petersen, S. D. Small, B. J. Sweitzer and L. L. Leape (1997). "The costs of adverse drug events in hospitalized patients." JAMA: the journal of the American Medical Association **277**(4): 307-311.

Belloni, A., V. Chernozhukov and Y. Wei (2013). "Honest confidence regions for a regression parameter in logistic regression with a large number of controls." arXiv preprint arXiv:1304.3969.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the Royal Statistical Society. Series B (Methodological): 289-300.

Berlin, J. A., S. C. Glasser and S. S. Ellenberg (2008). "Adverse event detection in drug development: recommendations and obligations beyond phase 3." American journal of public health **98**(8): 1366-1371.

Black, N. (1996). "Why we need observational studies to evaluate the effectiveness of health care." Bmj **312**(7040): 1215-1218.

Blum, M. D., D. J. Graham and C. A. McCloskey (1994). "Temafloxacin syndrome: review of 95 cases." Clinical Infectious Diseases **18**(6): 946-950.

Bodenreider, O. (2004). "The unified medical language system (UMLS): integrating biomedical terminology." Nucleic acids research **32**(suppl 1): D267-D270.

Brajovic, S., T. Piazza‐Hepp, L. Swartz and G. Dal Pan (2012). "Quality assessment of spontaneous triggered adverse event reports received by the Food and Drug Administration." Pharmacoepidemiology and drug safety **21**(6): 565-570.

Brewer, T. and G. A. Colditz (1999). "Postmarketing surveillance and adverse drug reactions: current perspectives and future needs." Jama **281**(9): 824-829.

Brookhart, M. A., T. Stürmer, R. J. Glynn, J. Rassen and S. Schneeweiss (2010). "Confounding control in healthcare database research: challenges and potential approaches." Medical care **48**(6): S114.

Campillos, M., M. Kuhn, A.-C. Gavin, L. J. Jensen and P. Bork (2008). "Drug target identification using side-effect similarity." Science **321**(5886): 263-266.

Cao, H., G. Hripcsak and M. Markatou (2007). "A statistical methodology for analyzing cooccurrence data from a large sample." Journal of biomedical informatics **40**(3): 343.

Caster, O. (2007). "Mining the WHO Drug Safety Database Using Lasso Logistic Regression." UUDM Project Report **16**.

Caster, O., G. N. Norén, D. Madigan and A. Bate (2010). "Large‐scale regression‐based pattern discovery: The example of screening the WHO global drug safety database." Statistical analysis and data mining **3**(4): 197-208.

Chapman, W. W., W. Bridewell, P. Hanbury, G. F. Cooper and B. G. Buchanan (2001). "A simple algorithm for identifying negated findings and diseases in discharge summaries." Journal of biomedical informatics **34**(5): 301-310.

Chen, E. S., G. Hripcsak, H. Xu, M. Markatou and C. Friedman (2008). "Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study." J Am Med Inform Assoc **15**(1): 87-98.

Chen, L. and C. Friedman (2004). "Extracting phenotypic information from the literature via natural language processing." Medinfo **11**(Pt 2): 758-762.

Classen, D. C., R. Resar, F. Griffin, F. Federico, T. Frankel, N. Kimmel, J. C. Whittington, A. Frankel, A. Seger and B. C. James (2011). "'Global trigger tool'shows that adverse events in hospitals may be ten times greater than previously measured." Health Affairs **30**(4): 581-589.

Coloma, P. (2012). Mining Electronic Healthcare Record Databases to Augment Drug Safety Surveillance, Erasmus MC: University Medical Center Rotterdam.

Coloma, P. M., P. Avillach, F. Salvo, M. J. Schuemie, C. Ferrajolo, A. Pariente, A. Fourrier-Réglat, M. Molokhia, V. Patadia and J. van der Lei (2013). "A reference standard for evaluation of methods for drug safety signal detection using electronic healthcare record databases." Drug safety **36**(1): 13-23.

Coloma, P. M., M. J. Schuemie, G. Trifirò, R. Gini, R. Herings, J. Hippisley‐Cox, G. Mazzaglia, C. Giaquinto, G. Corrao and L. Pedersen (2011). "Combining electronic healthcare databases in Europe to allow for large‐scale drug safety monitoring: the EU‐ADR Project." Pharmacoepidemiology and drug safety **20**(1): 1-11.

Coloma, P. M., G. Trifirò, V. Patadia and M. Sturkenboom (2013). "Postmarketing safety surveillance." Drug safety **36**(3): 183-197.

Coulter, D. and I. Edwards (1987). "Cough associated with captopril and enalapril." BMJ **294**(6586): 1521-1523.

Coulter, D. M. (2000). "The New Zealand intensive medicines monitoring programme in pro‐active safety surveillance." Pharmacoepidemiology and Drug Safety **9**(4): 273-280.

Cox, E., B. C. Martin, T. Van Staa, E. Garbe, U. Siebert and M. L. Johnson (2009). "Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: the International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report—Part II." Value in Health **12**(8): 1053-1061.

Curtis, J. R., H. Cheng, E. Delzell, D. Fram, M. Kilgore, K. Saag, H. Yun and W. DuMouchel (2008). "Adaptation of Bayesian data mining algorithms to longitudinal claims data: coxib safety as an example." Medical care **46**(9): 969.

DeLong, E. R., D. M. DeLong and D. L. Clarke-Pearson (1988). "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach." Biometrics: 837-845.

Duke, J. D., X. Han, Z. Wang, A. Subhadarshini, S. D. Karnik, X. Li, S. D. Hall, Y. Jin, J. T. Callaghan and M. J. Overhage (2012). "Literature based drug interaction prediction with clinical assessment using electronic medical records: novel myopathy associated drug interactions." PLoS computational biology **8**(8): e1002614.

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.

Eberth, J. M., K. N. Kline, D. A. Moskowitz, J. R. Montealegre and M. E. Scheurer (2014). "The Role of media and the internet on vaccine adverse event reporting: a case study of human papillomavirus vaccination." Journal of Adolescent Health **54**(3): 289-295.

Fawcett, T. (2006). "An introduction to ROC analysis." Pattern recognition letters **27**(8): 861-874.

Frazier, H. and G. Colditz (1995). "Oral contraceptives: post-marketing surveillance and rare, late complications of drugs." Medicine Worth Paying For: Assessing Medical Innovations. Harvard University Press, Cambridge, MA.

Freifeld, C. C., J. S. Brownstein, C. M. Menone, W. Bao, R. Filice, T. Kass-Hout and N. Dasgupta (2014). "Digital drug safety surveillance: monitoring pharmaceutical products in Twitter." Drug Safety **37**(5): 343-350.

Friedman, C. and N. Elhadad (2014). Natural language processing in health care and biomedicine. Biomedical Informatics, Springer**:** 255-284.

Friedman, C., T. C. Rindflesch and M. Corn (2013). "Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine." Journal of biomedical informatics **46**(5): 765-773.

Friedman, C., L. Shagina, Y. Lussier and G. Hripcsak (2004). "Automated encoding of clinical documents based on natural language processing." Journal of the American Medical Informatics Association **11**(5): 392-402.

Gagne, J. J., R. J. Glynn, J. A. Rassen, A. M. Walker, G. W. Daniel, G. Sridhar and S. Schneeweiss (2012). "Active safety monitoring of newly marketed medications in a distributed data network: application of a semi-automated monitoring system." Clinical Pharmacology & Therapeutics **92**(1): 80-86.

Gourieroux, C., A. Holly and A. Monfort (1982). "Likelihood ratio test, Wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters." Econometrica: Journal of the Econometric Society: 63-80.

Greenland, S. and H. Morgenstern (2001). "Confounding in health research." Annual Review of Public Health **22**(1): 189-212.

Grodstein, F., M. J. Stampfer, G. A. Colditz, W. C. Willett, J. E. Manson, M. Joffe, B. Rosner, C. Fuchs, S. E. Hankinson and D. J. Hunter (1997). "Postmenopausal hormone therapy and mortality." New England Journal of Medicine **336**(25): 1769-1776.

Gurwitz, J. H., T. S. Field, J. Avorn, D. McCormick, S. Jain, M. Eckler, M. Benser, A. C. Edmondson and D. W. Bates (2000). "Incidence and preventability of adverse drug events in nursing homes." The American journal of medicine **109**(2): 87-94.

Haerian, K., D. Varn, S. Vaidya, L. Ena, H. Chase and C. Friedman (2012). "Detection of pharmacovigilance-related adverse events using electronic health records and automated methods." Clinical Pharmacology & Therapeutics.

Hankinton, S. E., G. A. Colditz, D. J. Hunter, T. L. Spencer, B. Rosner and M. J. Stampfer (1992). "A quantitative assessment of oral contraceptive use and risk of ovarian cancer." Obstetrics & Gynecology **80**(4): 708-714.

Harpaz, R., W. DuMouchel, P. LePendu, A. Bauer-Mehren, P. Ryan and N. H. Shah (2013). "Performance of Pharmacovigilance Signal-Detection Algorithms for the FDA Adverse Event Reporting System." Clinical Pharmacology & Therapeutics.

Harpaz, R., W. DuMouchel, P. LePendu and N. H. Shah (2013). Empirical bayes model to combine signals of adverse drug reactions. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.

Harpaz, R., W. DuMouchel, N. Shah, D. Madigan, P. Ryan and C. Friedman (2012). "Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis." Clinical Pharmacology & Therapeutics **91**(6): 1010-1021.

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

Harpaz, R., D. Odgers, G. Gaskin, W. DuMouchel, R. Winnenburg, O. Bodenreider, A. Ripple, A. Szarfman, A. Sorbello and E. Horvitz (2014). "A time-indexed reference standard of adverse drug reactions." Scientific Data **1**.

Harpaz, R., S. Vilar, W. DuMouchel, H. Salmasian, K. Haerian, N. H. Shah, H. S. Chase and C. Friedman (2012). "Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions." Journal of the American Medical Informatics Association.

Harpaz, R., S. Vilar, W. DuMouchel, H. Salmasian, K. Haerian, N. H. Shah, H. S. Chase and C. Friedman (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.

Hauben, M. and J. K. Aronson (2009). "Defining 'signal'and its subtypes in pharmacovigilance based on a systematic review of previous definitions." Drug Safety **32**(2): 99-110.

Hauben, M., D. Madigan, C. M. Gerrits, L. Walsh and E. P. Van Puijenbroek (2005). "The role of data mining in pharmacovigilance."

Heeley, E., L. V. Wilton and S. A. Shakir (2002). "Automated signal generation in prescription-event monitoring." Drug Safety **25**(6): 423-432.

Hersh, W. R. (2007). "Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance." Am J Manag Care **81**: 126-128.

Hochberg, A. M., M. Hauben, R. K. Pearson, D. J. O'Hara, S. J. Reisinger, D. I. Goldsmith, A. L. Gould and D. Madigan (2009). "An evaluation of three signal-detection algorithms using a highly inclusive reference event database." Drug Safety **32**(6): 509-525.

http://www.nlm.nih.gov/research/umls/.

Huang, J. and C. X. Ling (2005). "Using AUC and accuracy in evaluating learning algorithms." Knowledge and Data Engineering, IEEE Transactions on **17**(3): 299-310.

Jewell, N. P. (2003). Statistics for epidemiology, Chapman & Hall/CRC.

Johnson, J. A. and J. L. Bootman (1995). "Drug-related morbidity and mortality: a cost-of-illness model." Archives of internal medicine **155**(18): 1949-1956.

Jonquet, C., N. H. Shah and M. A. Musen (2009). "The open biomedical annotator." Summit on translational bioinformatics **2009**: 56.

Kessler, D. A., S. Natanblut, D. Kennedy, E. Lazar, P. Rheinstein, C. Anello, D. Barash, I. Bernstein, R. Bolger and K. Cook (1993). "Introducing MEDWatch: a new approach to reporting medication and device adverse effects and product problems." JAMA **269**(21): 2765-2768.

Kohn, L. T., J. M. Corrigan and M. S. Donaldson (2000). To Err Is Human:: Building a Safer Health System, National Academies Press.

Kuhn, M., C. von Mering, M. Campillos, L. J. Jensen and P. Bork (2008). "STITCH: interaction networks of chemicals and proteins." Nucleic acids research **36**(suppl 1): D684-D688.

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

Leaman, R., L. Wojtulewicz, R. Sullivan, A. Skariah, J. Yang and G. Gonzalez (2010). Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks. Proceedings of the 2010 workshop on biomedical natural language processing, Association for Computational Linguistics.

LePendu, P., S. V. Iyer, A. Bauer-Mehren, R. Harpaz, J. M. Mortensen, T. Podchiyska, T. A. Ferris and N. H. Shah (2013). "Pharmacovigilance using clinical notes." Clinical pharmacology & therapeutics **93**(6): 547-555.

LePendu, P., S. V. Iyer, A. Bauer-Mehren, R. Harpaz, J. M. Mortensen, T. Podchiyska, T. A. Ferris and N. H. Shah (2013). "Pharmacovigilance Using Clinical Notes." Clinical Pharmacology & Therapeutics.

Li, Y., H. Salmasian, S. Vilar, H. Chase, C. Friedman and Y. Wei (2013). "A method for controlling complex confounding effects in the detection of adverse drug reactions using electronic health records." Journal of the American Medical Informatics Association.

Linder, J. A., J. S. Haas, A. Iyer, M. A. Labuzetta, M. Ibara, M. Celeste, G. Getty and D. W. Bates (2010). "Secondary use of electronic health record data: spontaneous triggered adverse drug event reporting." Pharmacoepidemiology and drug safety **19**(12): 1211-1215.

Lindquist, M., M. Ståhl, A. Bate, I. R. Edwards and R. H. Meyboom (2000). "A retrospective evaluation of a data mining approach to aid finding new adverse drug reaction signals in the WHO international database." Drug Safety **23**(6): 533-542.

Liu, L. and M. T. Zsu (2009). Encyclopedia of database systems, Springer Publishing Company, Incorporated.

Liu, M., Y. Wu, Y. Chen, J. Sun, Z. Zhao, X.-w. Chen, M. E. Matheny and H. Xu (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, S., W. Ma, R. Moore, V. Ganesan and S. Nelson (2005). "RxNorm: prescription for electronic drug information exchange."

Liu S, M. W., Moore R, Ganesan V, Nelson S (2005). "RxNorm: prescription for electronic drug information exchange " IT Professional **7**((5)): 17-23.

Makuch, R. W. (2006). "Detecting rare adverse events in postmarketing studies: sample size considerations." Drug information journal **40**(1): 89-98.

Manning, C. D., P. Raghavan and H. Schütze (2008). Introduction to information retrieval, Cambridge University Press Cambridge.

Medawar, C., A. Herxheimer, A. Bell and S. Jofre (2002). "Paroxetine, Panorama and user reporting of ADRs: Consumer intelligence matters in clinical practice and post-marketing drug surveillance." The International Journal of Risk and Safety in Medicine **15**(3): 161-169.

Meystre, S. M., G. K. Savova, K. C. Kipper-Schuler and J. Hurdle (2008). "Extracting information from textual documents in the electronic health record: a review of recent research." Yearb Med Inform **35**: 128-144.

Mitchell, A. S., D. A. Henry, R. Sanson-Fisher and D. L. O'Connell (1988). "Patients as a direct source of information on adverse drug reactions." Bmj **297**(6653): 891-893.

Nadkarni, P. M. (2010). "Drug safety surveillance using de-identified EMR and claims data: issues and challenges." Journal of the American Medical Informatics Association **17**(6): 671-674.

Nadkarni, P. M., L. Ohno-Machado and W. W. Chapman (2011). "Natural language processing: an introduction." Journal of the American Medical Informatics Association **18**(5): 544-551.

Neal, R. C., K. C. Ferdinand, J. Yčas and E. Miller (2009). "Relationship of ethnic origin, gender, and age to blood creatine kinase levels." The American Journal of Medicine **122**(1): 73-78.

Nikfarjam, A. and G. H. Gonzalez (2011). Pattern mining for extraction of mentions of adverse drug reactions from user comments. AMIA Annual Symposium Proceedings, American Medical Informatics Association.

Niu, M. T., D. E. Erwin and M. M. Braun (2001). "Data mining in the US Vaccine Adverse Event Reporting System (VAERS): early detection of intussusception and other events after rotavirus vaccination." Vaccine **19**(32): 4627-4634.

Norén, G. N., J. Hopstadius and A. Bate (2013). "Shrinkage observed-to-expected ratios for robust and transparent large-scale pattern discovery." Statistical methods in medical research **22**(1): 57-69.

Olsson, S. (1998). "The role of the WHO programme on international drug monitoring in coordinating worldwide drug safety efforts." Drug Safety **19**(1): 1-10.

Patadia, V. K., M. J. Schuemie, P. Coloma, R. Herings, J. van der Lei, S. Straus, M. Sturkenboom and G. Trifirò (2014). "Evaluating performance of electronic healthcare records and spontaneous reporting data in drug safety signal detection." International journal of clinical pharmacy: 1-11.

Patrick, A. R., S. Schneeweiss, M. A. Brookhart, R. J. Glynn, K. J. Rothman, J. Avorn and T. Stürmer (2011). "The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration." Pharmacoepidemiology and drug safety **20**(6): 551-559.

Platt, R., M. Wilson, K. A. Chan, J. S. Benner, J. Marchibroda and M. McClellan (2009). "The new Sentinel Network—improving the evidence of medical-product safety." New England Journal of Medicine **361**(7): 645-647.

Psaty, B. M. and S. P. Burke (2006). "Protecting the health of the public—Institute of Medicine recommendations on drug safety." New England Journal of Medicine **355**(17): 1753-1755.

Ramirez, E., A. Carcas, A. Borobia, S. Lei, E. Piñana, S. Fudio and J. Frias (2009). "A pharmacovigilance program from laboratory signals for the detection and reporting of serious adverse drug reactions in hospitalized patients." Clinical Pharmacology & Therapeutics **87**(1): 74-86.

Raschke, R. A., B. Gollihare, T. A. Wunderlich, J. R. Guidry, A. I. Leibowitz, J. C. Peirce, L. Lemelson, M. A. Heisler and C. Susong (1998). "A computer alert system to prevent injury from adverse drug events: development and evaluation in a community teaching hospital." Jama **280**(15): 1317-1320.

Rawson, N. S., G. L. Pearce and W. H. Inman (1990). "Prescription-event monitoring: methodology and recent progress." Journal of clinical epidemiology **43**(5): 509-522.

Reich, C. G., P. B. Ryan and M. J. Schuemie (2013). "Alternative outcome definitions and their effect on the performance of methods for observational outcome studies." Drug safety **36**(1): 181-193.

Rosenbaum, P. R. and D. B. Rubin (1983). "The central role of the propensity score in observational studies for causal effects." Biometrika **70**(1): 41-55.

Rothman, K. J., S. Greenland and T. L. Lash (2008). Modern epidemiology, Lippincott Williams & Wilkins.

Rothman, K. J., S. Lanes and S. T. Sacks (2004). "The reporting odds ratio and its advantages over the proportional reporting ratio." Pharmacoepidemiology and drug safety **13**(8): 519-523.

Ryan, P. B., D. Madigan, P. E. Stang, J. Marc Overhage, J. A. Racoosin and A. G. Hartzema (2012). "Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership." Statistics in medicine **31**(30): 4401-4415.

Ryan, P. B., D. Madigan, P. E. Stang, J. Marc Overhage, J. A. Racoosin and A. G. Hartzema (2012). "Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership." Statistics in Medicine.

Ryan, P. B., M. J. Schuemie, E. Welebob, J. Duke, S. Valentine and A. G. Hartzema (2013). "Defining a reference set to support methodological research in drug safety." Drug safety **36**(1): 33-47.

Ryan, P. B., P. E. Stang, J. M. Overhage, M. A. Suchard, A. G. Hartzema, W. DuMouchel, C. G. Reich, M. J. Schuemie and D. Madigan (2013). "A comparison of the empirical performance of methods for a risk identification system." Drug safety **36**(1): 143-158.

Safran, C., M. Bloomrosen, W. Hammond, S. Labkoff, S. Markel-Fox, P. C. Tang and D. E. Detmer (2007). "Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper." Journal of the American Medical Informatics Association **14**(1): 1-9.

Sakaeda, T., A. Tamon, K. Kadoyama and Y. Okuno (2013). "Data mining of the public version of the FDA Adverse Event Reporting System." International journal of medical sciences **10**(7): 796.

Sanchez-Guerrero, J., G. A. Colditz, E. W. Karlson, D. J. Hunter, F. E. Speizer and M. H. Liang (1995). "Silicone breast implants and the risk of connective-tissue diseases and symptoms." New England Journal of Medicine **332**(25): 1666-1670.

Santhi Swaroop Vege, M. "Etiology of acute pancreatitis. In: UpToDate, Basow, DS (Ed), UpToDate, Waltham, MA, 2012.".

Sarker, A., R. Ginn, A. Nikfarjam, K. O'Connor, K. Smith, S. Jayaraman, T. Upadhaya and G. Gonzalez (2015). "Utilizing social media data for pharmacovigilance: A review." Journal of biomedical informatics **54**: 202-212.

Savova, G. K., J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler and C. G. Chute (2010). "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES):

architecture, component evaluation and applications." Journal of the American Medical Informatics Association **17**(5): 507-513.

Schneeweiss, S. and J. Avorn (2005). "A review of uses of health care utilization databases for epidemiologic research on therapeutics." Journal of clinical epidemiology **58**(4): 323-337.

Schneeweiss, S., J. A. Rassen, R. J. Glynn, J. Avorn, H. Mogun and M. A. Brookhart (2009). "High-dimensional propensity score adjustment in studies of treatment effects using health care claims data." Epidemiology (Cambridge, Mass.) **20**(4): 512.

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.

Shang, N., H. Xu, T. C. Rindflesch and T. Cohen (2014). "Identifying plausible adverse drug reactions using knowledge extracted from the literature." Journal of biomedical informatics **52**: 293-310.

Shetty, K. D. and S. R. Dalal (2011). "Using information mining of the medical literature to improve drug safety." Journal of the American Medical Informatics Association: amiajnl-2011-000096.

Simpson, S. E. (2011). Self-controlled methods for postmarketing drug safety surveillance in large-scale longitudinal data, Columbia University.

Stang, P. E., P. B. Ryan, J. A. Racoosin, J. M. Overhage, A. G. Hartzema, C. Reich, E. Welebob, T. Scarnecchia and J. Woodcock (2010). "Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership." Annals of internal medicine **153**(9): 600.

Stratton, K., A. Baciu and S. P. Burke (2007). The Future of Drug Safety:: Promoting and Protecting the Health of the Public, National Academies Press.

Strom, B. L. (2011). "Sample size considerations for pharmacoepidemiologic studies." Pharmacoepidemiology: 52.

Stürmer, T., M. Joshi, R. J. Glynn, J. Avorn, K. J. Rothman and S. Schneeweiss (2006). "A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods." Journal of clinical epidemiology **59**(5): 437. e431-437. e424.

Tatonetti, N., J. Denny, S. Murphy, G. Fernald, G. Krishnan, V. Castro, P. Yue, P. Tsau, I. Kohane and D. Roden (2011). "Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels." Clinical Pharmacology & Therapeutics **90**(1): 133-142.

Tatonetti, N. P., G. H. Fernald and R. B. Altman (2012). "A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports." Journal of the American Medical Informatics Association **19**(1): 79-85.

Tatonetti, N. P., P. Y. Patrick, R. Daneshjou and R. B. Altman (2012). "Data-driven prediction of drug effects and interactions." Science translational medicine **4**(125): 125ra131.

Tatonetti, N. P., P. P. Ye, R. Daneshjou and R. B. Altman (2012). "Data-driven prediction of drug effects and interactions." Sci Transl Med **4**: 125ra131.

Thomas, E. J., D. M. Studdert, J. P. Newhouse, B. I. Zbar, K. M. Howard, E. J. Williams and T. A. Brennan (1999). "Costs of medical injuries in Utah and Colorado." Inquiry: 255-264.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological): 267-288.

Tisdale, J. E. and A. Douglas (2010). Drug-induced diseases: prevention, detection, and management, ASHP.

Topol, E. J. (2004). "Failing the public health—rofecoxib, Merck, and the FDA." New England Journal of Medicine **351**(17): 1707-1709.

Trifiro, G., A. Fourrier-Reglat, M. Sturkenboom, C. D. Acedo and J. Van Der Lei (2009). "The EU-ADR project: preliminary results and perspective." Stud Health Technol Inform **148**: 43-49.

Trifirò, G., A. Pariente, P. M. Coloma, J. A. Kors, G. Polimeni, G. Miremont‐Salamé, M. A. Catania, F. Salvo, A. David and N. Moore (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.

Upton, G. J. G. (1992). "Fisher's exact test." Journal of the Royal Statistical Society. Series A (Statistics in society): 395-402.

van Puijenbroek, E. P., A. Bate, H. G. Leufkens, M. Lindquist, R. Orre and A. C. Egberts (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.

Vilar, S., R. Harpaz, H. S. Chase, S. Costanzi, R. Rabadan and C. Friedman (2011). "Facilitating adverse drug event detection in pharmacovigilance databases using molecular structure similarity: application to rhabdomyolysis." Journal of the American Medical Informatics Association **18**(Suppl 1): i73-i80.

Vilar, S., P. Ryan, D. Madigan, P. Stang, M. Schuemie, C. Friedman, N. Tatonetti and G. Hripcsak (2014). "Similarity-Based Modeling Applied to Signal Detection in Pharmacovigilance." CPT: pharmacometrics & systems pharmacology **3**(9): e137.

Wang, H.-w., A. M. Hochberg, R. K. Pearson and M. Hauben (2010). "An experimental investigation of masking in the US FDA adverse event reporting system database." Drug safety **33**(12): 1117-1133.

Wang, W., K. Haerian, H. Salmasian, R. Harpaz, H. Chase and C. Friedman (2011). A drug-adverse event extraction algorithm to support pharmacovigilance knowledge mining from PubMed citations. AMIA Annual Symposium Proceedings, American Medical Informatics Association.

Wang, X., G. Hripcsak and C. Friedman (2009). "Characterizing environmental and phenotypic associations using information theory and electronic health records." BMC bioinformatics **10**(Suppl 9): S13.

Wang, X., G. Hripcsak, M. Markatou and C. Friedman (2009). "Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study." Journal of the American Medical Informatics Association **16**(3): 328-337.

Warrian, W. G., J. A. Halikas, R. D. Crosby, G. A. Carlson and F. Crea (1992). "Observations on increased CPK levels in" asymptomatic" cocaine abusers." Journal of addictive diseases **11**(4): 83-95.

White, R. W., R. Harpaz, N. H. Shah, W. DuMouchel and E. Horvitz (2014). "Toward enhanced pharmacovigilance using patient-generated data on the internet." Clinical Pharmacology & Therapeutics **96**(2): 239-246.

White, R. W., N. P. Tatonetti, N. H. Shah, R. B. Altman and E. Horvitz (2013). "Web-scale pharmacovigilance: listening to signals from the crowd." Journal of the American Medical Informatics Association: amiajnl-2012-001482.

Wilson, A. M., L. Thabane and A. Holbrook (2004). "Application of data mining techniques in pharmacovigilance." British journal of clinical pharmacology **57**(2): 127-134.

Wysowski, D. K. and L. Swartz (2005). "Adverse drug event surveillance and drug withdrawals in the United States, 1969-2002: the importance of reporting suspected reactions." Archives of internal medicine **165**(12): 1363-1369.

Xu, H., S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman and J. C. Denny (2010). "MedEx: a medication information extraction system for clinical narratives." Journal of the American Medical Informatics Association **17**(1): 19-24.

Xu, R. and Q. Wang (2014). "Large-scale combining signals from both biomedical literature and the FDA Adverse Event Reporting System (FAERS) to improve post-marketing drug safety signal detection." BMC bioinformatics **15**(1): 17.

Zorych, I., D. Madigan, P. Ryan and A. Bate (2013). "Disproportionality methods for pharmacovigilance in longitudinal observational databases." Statistical methods in medical research **22**(1): 39-56.

Zou, H. and T. Hastie (2005). "Regularization and variable selection via the elastic net." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67**(2): 301-320.