Integrating Domain Knowledge to Improve Signal Detection from Electronic Health Records for Pharmacovigilance

Ning Shang

Follow this and additional works at: http://digitalcommons.library.tmc.edu/uthshis_dissertations
Part of the Bioinformatics Commons, and the Medicine and Health Sciences Commons

Recommended Citation
http://digitalcommons.library.tmc.edu/uthshis_dissertations/26

This is brought to you for free and open access by the School of Biomedical Informatics at DigitalCommons@TMC. It has been accepted for inclusion in UT SBMI Dissertations (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact laurel.sanders@library.tmc.edu.
Integrating Domain Knowledge to Improve Signal Detection from Electronic Health Records for Pharmacovigilance

By

Ning Shang, MS

APPROVED:

Trevor Cohen, MBChB, PhD, Chair

Jorge Herskovic, MD, PhD

Hua Xu, PhD

Elmer V. Bernstam, MD, MSE, MS

Peng Wei, PhD

Date approved: 7/14/2014
Integrating Domain Knowledge to Improve Signal Detection from Electronic Health Records for Pharmacovigilance

A Dissertation

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
School of Biomedical Informatics
in Partial Fulfilment of the Requirements for the Degree of
Doctor of Philosophy

By

Ning Shang, MS

University of Texas Health Science Center at Houston
2014

Dissertation Committee:

Trevor Cohen, MBChB, PhD\(^1\), Chair
Jorge Herskovic, MD, PhD\(^2\)
Hua Xu, PhD\(^1\)
Elmer V. Bernstam, MD, MSE, MS\(^1\)
Peng Wei, PhD\(^3\)
\(^1\)The School of Biomedical Informatics
\(^2\)MD Anderson Cancer Center
\(^3\)The School of Public Health
Copyright by
Ning Shang
2014
Acknowledgements

I want to thank Dr. Trevor Cohen for being a great mentor. His scientific spirit, serious attitude and considerate character have taught me to become a thorough scientist and beyond. I want to thank Dr. Jorge Herskovic for his support and all the enlightening discussions I had with him. I want to thank Dr. Hua Xu for sharing his research experiences with me. I want to thank Dr. Elmer V. Bernstam and his team for building and sharing clinical data repositories, facilitating my research goals. I want to thank Dr. Peng Wei for providing invaluable advice for selecting the best statistical models. Last but not least, I want to thank Dr. Parsa Mirhaji for opening the world of pharmacovigilance to me and the training in out-of-the-box thinking.

I want to thank my parents, my sister, and my fiancée for their unconditional love and support.

This work was supported by the US National Library of Medicine Grant (1R01LM011563), Using Biomedical Knowledge to Identify Plausible Signals for Pharmacovigilance.
Abstract

The intent of this dissertation is to make a contribution to the field of pharmacovigilance. Pharmacovigilance, also known as post-marketing drug surveillance, is the process of continued monitoring for adverse drug reactions (ADRs) after drugs are released into the market. An ADR is a harmful or unpleasant reaction related to the use of a medical product. ADRs were reported to be between the fourth and sixth leading cause of death in the United States in 1994, accounting for 3-7% of medical hospital admissions. On account of the practice of pharmacovigilance, Vioxx (Rofecoxib) and Avandia (Rosiglitazone) are examples of high profile drugs that were suspended from the American or European market.

To prevent these effects on human health, pre-marketing clinical trials are designed to test drug safety and efficacy. Although clinical trials are extensive and last multiple years, rare ADRs may not be detected, and others may occur on account of idiosyncratic characteristics of individuals excluded from the evaluated sample.

To aid the pharmacovigilance process, automated methods for the identification of strongly correlated drug/ADR pairs from data sources such as adverse event reporting systems, or Electronic Health Records (EHRs), have been developed. These methods however are generally statistical in nature, and do not draw upon the large volumes of knowledge embedded in the biomedical literature.
In this dissertation I investigate the ability of scalable Literature Based Discovery (LBD) methods to identify side effects of pharmaceutical agents in a computationally automated manner. LBD methods can provide evidence from the literature to support the plausibility of a drug/ADR association, thereby assisting human review to validate the signal, which is an essential component of pharmacovigilance. The hypothesis underlying this work is that by combining signals mined from EHR data with biomedical domain knowledge, the accuracy of side effects detection may be improved. This also addresses the lack of causality assessment in existing statistical methods in pharmacovigilance practice.

My theoretical contribution is that by conducting automated abductive reasoning and by estimating the strength of generated explanatory hypotheses the plausibility of a drug/ADR signal can be assessed. I adapt and extend the original abductive reasoning process as defined by Peirce in 19th century by stating that the strength of the explanations found for an observation is a measure for its plausibility, rather than taking an observation as given.

Practical contributions to pharmacovigilance and informatics include the development of methods to leverage the knowledge from biomedical literature, the detection of signals from the EHR data and the subsequent evaluation using supporting evidence from the literature on a large scale in an automated way, and the development of an improved drug/ADR reference set. My contributions are not restricted to pharmacovigilance and as such constitute a contribution to the field of informatics in general.

I demonstrate that my work has extended the state of the art in EHR-based pharmacovigilance and contribute new ideas that pave the way for further studies with the potential to further enhance the field of pharmacovigilance and drug safety.
# Table of Contents

Acknowledgements................................................................................................................. ii

Abstract.................................................................................................................................. iii

Table of Contents....................................................................................................................... v

List of Tables ............................................................................................................................. xii

List of Figures ............................................................................................................................ xiii

Chapter 1: Introduction .............................................................................................................. 1

1. 1 Significance of Pharmacovigilance .................................................................................. 1

1.2 Challenges in pharmacovigilance and suggested solutions ............................................. 2

1.3 Dissertation structure ....................................................................................................... 4

1.4 Key contributions .............................................................................................................. 5

Chapter 2: Literature Review .................................................................................................... 7

2.1 Pharmacovigilance: post-marketing drug surveillance .................................................... 7

2.1.1 Pharmacovigilance definition and significance ............................................................ 7

2.1.2 Pharmacovigilance workflow in related health departments ....................................... 9

2.1.3 Pharmacovigilance data sources .................................................................................. 10

2.2 Pharmacovigilance methodology .................................................................................... 12

2.2.1 Statistical data mining from pharmacovigilance database ........................................... 12
2.9.2 Using distributional semantic models to find plausible drug/ADR associations

Chapter 3: Chapter 3: Signal Detection from Outpatient Electronic Health Record

Data Using Statistical Mining

3.1 Materials

3.1.1 Clinical data warehouse (CDW)

3.1.2 Medical language extraction and encoding system (MedLEE)

3.1.3 Side effect resource 2 (SIDER2)

3.1.4 Hierarchical relationships between concepts from UMLS knowledge base

3.1.5 MEDication-Indication (MEDI)

3.2 Methods

3.2.1 Annotating the EHR data using MedLEE

3.2.2 Construct test data set

3.2.3 Retrieve co-occurrence data for SIDER2 test set

3.2.4 Statistical algorithms

3.2.4.1 Disproportionality analysis

3.2.4.2 Other statistical algorithm based on co-occurrence

3.2.5 Filtering indications

3.2.6 Performance evaluation

3.3 Experimental design
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4</td>
<td>Results</td>
<td>45</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Experiment dataset</td>
<td>45</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Query the co-occurrence data for experiment datasets</td>
<td>46</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Statistical results</td>
<td>47</td>
</tr>
<tr>
<td>3.4.3.1</td>
<td>Performance results</td>
<td>47</td>
</tr>
<tr>
<td>3.4.3.1.1</td>
<td>Performance of statistical algorithms with original co-occurrence data</td>
<td>47</td>
</tr>
<tr>
<td>3.4.3.1.2</td>
<td>Effects of potential ontology intervention</td>
<td>48</td>
</tr>
<tr>
<td>3.4.3.1.3</td>
<td>Effects of MEDI intervention</td>
<td>50</td>
</tr>
<tr>
<td>3.5</td>
<td>Discussion</td>
<td>52</td>
</tr>
<tr>
<td>3.6</td>
<td>Conclusion</td>
<td>54</td>
</tr>
</tbody>
</table>

Chapter 4: Using Knowledge Extracted from the Literature to Identify Plausible Adverse Drug Reactions

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Materials</td>
<td>55</td>
</tr>
<tr>
<td>4.1.1</td>
<td>MMB and SemMedDB</td>
<td>56</td>
</tr>
<tr>
<td>4.1.2</td>
<td>SIDER2</td>
<td>56</td>
</tr>
<tr>
<td>4.1.3</td>
<td>MEDication-Indication (MEDI)</td>
<td>57</td>
</tr>
<tr>
<td>4.2</td>
<td>Methods</td>
<td>58</td>
</tr>
<tr>
<td>4.2.1</td>
<td>RRI</td>
<td>58</td>
</tr>
<tr>
<td>4.2.2</td>
<td>PSI</td>
<td>63</td>
</tr>
<tr>
<td>4.2.2.1</td>
<td>Operations in PSI</td>
<td>63</td>
</tr>
</tbody>
</table>
5.2 Methods

5.2.1 Import SIDER2 and FAERS data in database

5.2.2 Parsing drug and side effect terms using MetaMap

5.2.3 Mapping between SIDER2 and FAERS

5.2.4 Retrieve number of reports for mapped SIDER2 drug-side effect pairs

   and perform disproportionality analysis to find significant SIDER2 drug/ADR associations

5.2.5 Performance measure

5.3 Experimental design

5.4 Results

5.4.1 Concept extraction

5.4.2 Reported SIDER side effects

5.4.3 Statistically significant drug/ADR pairs detected by disproportionality measures

5.4.4 Manual review sample drug/ADR pairs without case reports

5.5 Discussion

5.6 Conclusion

Chapter 6: Improving Signal Detection from the Electronic Health Records by Using Literature Based Discovery

6.1 Experimental design
List of Tables

Table 2-1. Quantitative data mining procedures used as disproportionality measures in SRS ................................................................. 15
Table 3-1. Overall error measure in FDR approach ................................................................. 43
Table 3-2. Top 10 clinical documentation types in the EHR system ........................................ 46
Table 3-3. Number of false pairs that have been excluded as indications ................................ 51
Table 4-1. Different groups by extending MEDI in experiment 2 ........................................ 72
Table 4-2. The most frequent predicate paths in inferred discovery patterns .............................. 74
Table 4-3. Results of precision and rank-based measures for different groups .............................. 76
Table 4-4. PSI-double+triple model performance across all tests with different MEDI lists .................. 80
Table 4-5. Reasoning pathways used to retrieve evidence from the literature for the pair rosiglitazone -- myocardial infarction ........................ 82
Table 4-6. Some example predications for possible mechanism of rosiglitazone causing myocardial infarction ...................................... 83
Table 5-1. The percentage of statistically significant SIDER2 drug/ADR pairs using disproportionality measures ........................................ 109
Table 6-1. SIDER2 and a Reference Set are used to construct the dataset for the experiment ............... 115
Table 6-2. Results of precision and rank-based measures for different groups and different datasets ...................................................... 117
List of Figures

Figure 2-1. Number of adverse event reports reported to FDA from 2001 to 2011 .................................................................9
Figure 2-2. FDA pharmacovigilance process ........................................11
Figure 3-1. The example ancestor-offspring hierarchy relationships between UMLS concepts .........................................................39
Figure 3-2. Comparison of ability of different performance metrics to recover SIDER2 side effects for different statistical algorithms with different drug/ADR co-occurrence threshold ........................................48
Figure 3-3. Comparison of the predicted drug/ADR association count comparing between original and descendant co-occurrence set ........49
Figure 3-4. Comparison of ability of different performance metrics to recover SIDER2 side effects for different statistical algorithms with different co-occurrence set ......................................................50
Figure 3-5. Comparison of the predicted drug/ADR association count comparing between different MEDI filtering indications ..........51
Figure 3-6. Comparison of ability of different performance metrics to recover SIDER2 side effects for different statistical algorithms after filtering indications by different MEDI lists .......................52
Figure 4-1. The pseudo code for RRI model training process ...............62
Figure 4-2. Schematic representation of RRI training and inference process ......62
Figure 4-3. The Pseudo code for PSI model training process ..................65
Figure 4-4. Schematic representation of PSI training and inference process ........68
Figure 4-5. Experimental design in the detection of SIDER2 known ADRs using LBD distributional semantic models .........................69
Figure 4-6. ROC plot of true positive rate and false positive rate for all groups .................................................................77
Figure 4-7. Performance of AUC for all drugs by PSI-double+triple group ............78
Figure 4-8. Rediscovery plot for experiment groups ........................................79
Figure 4-9. The predications retrieved by reasoning pathway for rosiglitazone
causing myocardial infarction with specifying semantic
groups for concepts .................................................................85
Figure 4-10. Middle terms that were retrieved by PSI discovery patterns
involving LDL-C ........................................................................87
Figure 5-1. The procedure for mapping SIDER2 and FAERS .........................103
Figure 5-2. Procedure of mapping SIDER2 drugs/side effects to FAERS
drugs/reactions using “dipyridamole” as an example ................................104
Figure 5-3. Study design for decreasing false SIDER2
drug-side effect associations ..........................................................106
Figure 5-4. Frequency distribution of number of SIDER2 drug/ADR pairs
grouped by the number of FAERS reports that contain the
drug/ADR of interest ...................................................................108
Figure 5-5. Interpretation of 60 randomly selected SIDER2 pairs without
any supporting reports .................................................................110
Figure 6-1. Overall research design for reranking statistically significant
drug/ADR associations by similarity scores ........................................114
Figure 6-2. Comparing accumulated number of true positives for each ranking
between pre- and post-reranking signals for
SIDER2 set and the Reference set ................................................118
Figure 6-3. ROC plot of true positive rate and false positive rate
for pre- and post-reranking groups for SIDER2 set .........................119
Figure 6-4. ROC plot of true positive rate and false positive rate
for pre- and post-reranking groups for Reference set ......................120
Figure 6-5. The proposed architecture for a plausibility-based
pharmacovigilance system ............................................................122
**Chapter 1: Introduction**

1.1 Significance of Pharmacovigilance

An adverse drug reaction (ADR) is an “appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medical product” (Edwards & Aronson, 2000). ADRs contribute to substantial morbidity, health care visits and hospital admissions (Baker et al., 2004; Chan, Nicklason, & Vial, 2001; Hamilton, Briceland, & Andritz, 1998; Lazarou, Pomeranz, & Corey, 1998; Pirmohamed et al., 2004). For example, ADRs were reported to be between the fourth and sixth leading cause of death in the United States in 1994 (Lazarou et al., 1998), accounting for 3-7% of medical hospital admissions (Baker et al., 2004; Hamilton et al., 1998) and a substantial number of health care visits (Bourgeois, Mandl, Valim, & Shannon, 2009). A recent observational study conducted in two emergency departments shows that 8.9% of emergency patients had a probable ADR (Capuano et al., 2004). Overall, they have a considerable negative impact and place an enormous burden on both health and healthcare system (Pirmohamed et al., 2004; Rodríguez-Monguíó, Otero, & Rovira, 2003) and consequently constitute a notable public health problem (Capuano et al., 2004).

To prevent these harmful effects on human health, pre-marketing clinical trials are designed to test drug safety and efficacy. Phase III clinical trials have been estimated to cost 86.3 million US dollars and last 30.5 months on average (DiMasi, Hansen, & Grabowski, 2003). Nonetheless, rare ADRs may not be detected due to the limited duration
and sample size of such trials, and others may occur on account of idiosyncratic characteristics of individuals excluded from the evaluated sample. The continued monitoring for ADRs after drugs are released into the market, referred to as pharmacovigilance (also known as post-marketing drug surveillance), is therefore an important tool to monitor and improve drug safety (Mann & Andrews, 2007). For example, on account of the practice of pharmacovigilance, Vioxx (Rofecoxib) was withdrawn voluntarily worldwide in 2004 because of the finding of increased risk of heart attack (Merck, 2004). Similarly, Avandia (Rosiglitazone) was suspended from the European market in 2010 by the European Medicines Agency (EMA) (FDA, 2011; Ye, 2011).

1.2 Challenges in pharmacovigilance and suggested solutions

Over the last decade, the World Health Organization (WHO), the Food and Drug Administration (FDA), EMA and others instituted spontaneous reporting systems (SRSs) for the systematic collection of ADR reports (Rawlins, 1988a). Drug safety data obtained from SRSs have been analyzed using quantitative data mining procedures to retrieve strongly associated drug/ADR pairs (Puijenbroek, Diemont, & van Grootheest, 2003; Rawlins, 1988a, 1988b). These highlighted associations are subsequently reviewed and scrutinized by domain experts. These are essential parts of a pharmacovigilance system, which are referred to as drug/ADR signal detection and evaluation (A. Bate et al., 1998; A. Bate, Lindquist, Edwards, & Orre, 2002; Bates, Lindquist M., Orre R., Edwards I., & Meyboom R., 2002).

There are two major problems with these processes. First, research suggests data collected by SRS are limited by long time latency, incorrect or incomplete clinical information,
underreporting and reporting bias (Rawlins, 1988a, 1988b; Hasford, Goettler, Munter, & Müller-Oerlinghausen, 2002; Alvarez-Requejo et al., 1998). Second, it has been argued that causality assessment is lacking in pharmacovigilance practice (Anderson & Borlak, 2011). The WHO defines causality assessment as the evaluation of the likelihood that a medicine caused the observed adverse event (WHO-Uppsala Monitoring Centre, 2013). The lack of this evaluation is due to the expense and time involved in manual review by domain experts. In practice resources are too limited for the large number of signals produced by SRSs.

A potential workaround to the resource limitations is to use information from other sources to supplement SRSs. Drug intake and symptoms are documented by clinicians in EHR data, although the rate of false positives is expected to be high because the number of possible adverse events is small compared to the number of desired treatment outcomes. After filtering known drug indications a causality assessment of statistically significant signals is necessary to identify possible side effects. Potential and known mechanisms of action are described in detail in the biomedical literature. It is therefore possible to leverage domain knowledge from biomedical literature to make statements about the plausibility of drug/ADR connections. The hypothesis underlying this work is that by combining signals mined from EHR data with biomedical domain knowledge, the accuracy of side effect detection may be improved.

I evaluate this hypothesis in a series of experiments. First, I apply disproportionality measures and Chi-square test to analyze drug-event co-occurrence data derived from EHR to find statistically significant drug/ADR associations and evaluate the performance of these statistical algorithms. For the statistically significant drug/ADR associations, I apply
scalable methods of literature-based discovery (LBD) (Hristovski, Friedman, Rindflesch, & Peterlin, 2006; Kostoff, Briggs, Solka, & Rushenberg, 2008; Swanson, 1986b) leveraging methods of distributional semantics (Cohen, Schvaneveldt, & Widdows, 2010; Cohen, Widdows, Schvaneveldt, Davies, & Rindflesch, 2012; Widdows & Cohen, 2010) to evaluate the plausibility of their connections. I do this by utilizing knowledge from the published literature to automatically generate explanatory hypotheses which collectively contribute to a plausibility score and a subsequent reranking according to their plausibility. This also alleviates the problem of an overwhelming amount of associations for manual expert review.

During the course of this research, it came to my attention that there is no agreed-upon gold standard in pharmacovigilance (Tatonetti, Ye, Daneshjou, & Altman, 2012). This may affect the performance evaluation because of the existing false drug/ADR associations in the reference data. So a reference set was developed and used for additional performance evaluation in this study.

To reiterate, the unifying hypothesis that is tested in the dissertation is that drug/ADR signal detection from EHRs can be improved by integrating knowledge from the biomedical literature.

1.3 Dissertation structure

This dissertation is structured as follows. In Chapter 2, I review literature related to the field of and recent research in pharmacovigilance as well as LBD methods and distributional semantics. In the following chapters I perform a series of experiments to evaluate my hypotheses. In Chapter 3, I describe the first experiment, in which statistical
mining algorithms are used to find statistically correlated drug/ADR associations from EHR data. In Chapter 4, I build LBD based distributional semantic models to identify plausible ADRs using knowledge extracted from the literature. In Chapter 5, a drug/ADR association reference set is built as a subset of a dataset containing known side effects. The subset contains only those associations that are confirmed to be statistically significant using data from an SRS. In Chapter 6, I evaluate the effects of plausibility based reranking on the precision of the statistically significant drug/ADR associations detected in EHR data. In this experiment, two reference sets are utilized and the performance against these sets is compared. The dissertation concludes with describing its significance and contributions to the field of pharmacovigilance, informatics and public health.

1.4 Key contributions

The purpose of this dissertation is to automatically identify drug/ADR associations that are most likely causal among those signals detected from EHR with statistical methods by integrating domain knowledge from the biomedical literature that is related to side effects.

- A surprising drug/ADR signal (surprise observation) is detected from an EHR system.
- There is a genuine doubt about this observation (doubt).
- To resolve the doubt, an inquiry is initiated (inquiry) by searching the literature to support plausible connections between the drug and the ADR.
- Supporting plausibility evidence is subsequently utilized as sufficient explanation (explanation) to relieve the doubt about the surprising observation.

My theoretical contribution is that by conducting automated abductive reasoning (Josephson & Josephson, 1996; Locke, Golden-Biddle, & Feldman, 2008; Peirce, 1955;
Schvaneveldt & Cohen, 2010) and by estimating the strength of generated explanatory hypotheses the plausibility of an observation (in this work, a drug/ADR signal) can be assessed. This means that I adapt the original abductive reasoning process by stating that the strength of the explanations found for an observation is a measure for its plausibility. A low plausibility likely implies that an observation is a false positive and should be considered for rejection. A high measure of plausibility on the contrary supports the validity of the observation.

This differs from the original line of thought by introducing a measure to assess plausibility rather than taking the observation as a given fact.

The practical contributions to pharmacovigilance and informatics are as follows:

- The development of methods to leverage knowledge from the literature as a means to assess the plausibility of observations from EHR data.
- The detection of signals from EHR data and the subsequent evaluation using supporting evidence from literature in pharmacovigilance on a large scale in an automated way.
- The development of a drug/ADR reference set based on both an established side effects repository and a high volume drug/ADR data set from an SRS.
- The methods and procedures of integrating formal knowledge can be generalized and applied to other domains, such as outbreak detection, for which a timely identification of plausibility for signals is essential. This is not restricted to pharmacovigilance and constitutes a contribution to the field of informatics in general.
Chapter 2: Literature Review

2.1 Pharmacovigilance: post-marketing drug surveillance

2.1.1 Pharmacovigilance definition and significance

Vioxx (Rofecoxib) was withdrawn voluntarily from market by Merck in 2004, after it was found that the use of this agent increased the risk of myocardial infarction (Merck, 2004). Graham et al (Graham et al., 2005) estimated that between 88,000 and 140,000 excess serious coronary heart diseases might have been caused by rofecoxib in the US since it was launched in 1999. Avandia (Rosiglitazone) was suspended from the European market in 2010 (Blind, Dunder, Graeff, & Abadie, 2011; EMA, 2010; Ye, 2011) on account of an increased risk of cardiovascular complications. These high-profile examples illustrate that pharmacovigilance is very important to supplement existing drug safety profiles because clinical drug trials cannot be large or long enough to identify all problems related to a new drug (Rawlins, 1988a). Additionally, subjects are pre-selected by eligibility criteria and therefore may not fully represent the patient population after the drugs are put to market (Anderson & Borlak, 2011). Besides, some patient may have idiosyncratic drug reactions which cannot be explained by the pharmacological effect of the drug (Knowles, Uetrecht, & Shear, 2000). Consequently, it is highly unlikely that instances of all possible ADRs will be detected during pre-marketing clinical trials.

To address this problem, health departments and organizations (such as the World Health Organization (WHO), United States Food and Drug Administration (FDA), and European
Medicines Agency (EMA)) encourage physicians, other health care professionals, and patients to report voluntarily about any observed ADRs. In addition to voluntarily reporting, pharmaceutical companies are mandatory required to report serious adverse events (Ahmad, 2003). These bodies have Spontaneous Reporting Systems (SRS) to enable the efficient submission of reports electronically (Kessler et al., 1993; Wysowski & Swartz, 2005).

Since 1969, the FDA Adverse Event Reporting System (FAERS), which provides the means for clinicians to report suspected adverse events electronically, collected more than 7 million case reports (Fine, 2013) and the number of ADR reports submitted to the FDA continues to grow (Figure 2-1). More than 75 drug products have been removed from the market due to safety problems from 1969 to 2002 (Wysowski & Swartz, 2005). The facts emphasize the importance of post-marketing drug monitoring, known as pharmacovigilance – “the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem after drugs are on market” (WHO, 2010). Pharmacovigilance is designed to detect any rare or long-term adverse effects over a very large population and a long period of time.
2.1.2 Pharmacovigilance workflow in related health departments

In general, the pharmacovigilance process proceeds as follows (Andrew Bate, 2003; Edwards, 1992; Lindquist et al., 1999):

1. Reported drug-related problems are collected in SRSs nationally or internationally;
2. Quantitative data mining procedures are used to analyze these data and retrieve relatively strongly correlated drug/ADR pairs (drug/ADR associations);
(3) These highlighted associations are then reviewed and evaluated by domain experts making up an expert clinical review panel; and

(4) Associations considered to be of clinical interest are then annotated as “signals”.

Specifically, signal is defined as “reported information on a possible causal relationship between an adverse event and a drug, the relationship being unknown or incompletely documented previously” (Edwards & Biriell, 1994; van Puijenbroek et al., 2002). Overall, the pharmacovigilance process in WHO (Andrew Bate, 2003), FDA (Fine, 2013; Gould, 2005) and EU-ADR initiative (P. Lopes et al., 2013) includes two components -- a statistical component (quantitative signal detection, steps (1) and (2)) and a qualitative component (expert clinical review, steps (3) and (4)) (Andrew Bate, 2003). Figure 2-2 demonstrates the specific pharmacovigilance workflow in FDA.

2.1.3 Pharmacovigilance data sources

Through pharmacovigilance, international and national health institutions gather large amount of data mainly from SRS for further analysis. In addition to case reports, specialized networks and active surveillance data are also considered as data sources for detecting safety signals, even though they are in experimental phrases (FDA Science Board Subcommittee, 2011). Specialized networks, such as the Drug Induced Liver Injury Network (DILIN) (Hoofnagle, 2004), focus on collecting side effects data concerning specific organ targets, patterns of ADRs or patient populations. Active surveillance efforts, such as Sentinel Network (Platt et al., 2009; Platt, Madre, Reynolds, & Tilson, 2008) and Observational Medical Outcomes Partnership (OMOP) (Stang et al., 2010), utilize existing systems for the purpose of drug monitoring (U.S. Food and Drug Administration, 2008). Examples of potentially useful systems include EHRs, pharmacoepidemiological databases
(Johansson, Wallander, de Abajo, & Rodriguez, 2010), health insurance claims databases (Brown et al., 2007, 2009), and so forth. the Observational Medical Outcomes Partnership (OMOP) uses EHR and other data sources to identify ADR signals and the preliminary results demonstrate that active surveillance from EHR systems can complement current pharmacovigilance practice, but also that performance varies by data source, drug and outcomes of interest (FDA Science Board Subcommittee, 2011).

Figure 2-2: FDA pharmacovigilance process – adapted from (Fine, 2013; Gould, 2005)

With the opportunity presented by EHRs’ broader availability, researchers have also used EHR data for drug safety research and have attempted to demonstrate that EHR data are a relevant and significant data source for pharmacovigilance (Collins, 2011; Wang,
These authors argue that EHR data can compensate for some of the deficiencies of SRS, such as under-reporting, misclassification, a long lag time between observation and reporting, reporting bias and the provision of incomplete background clinical information (Rawlins, 1988a, 1988b).

2.2 Pharmacovigilance methodology

2.2.1 Statistical data mining from pharmacovigilance database

Statistical algorithms 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. Quantitative data mining procedures are measuring “disproportionality”. Reporting disproportionality is defined as a statistically significant higher reporting frequency of particular drug/ADR associations compared with marginal distributions of drugs and ADRs as a background in the reporting database (Shibata & Hauben, 2011). These disproportionality measures quantify how often a drug and a possible event co-occur compared to the background reporting occurrence across all other drugs and events using an independence model (Manfred Hauben, Madigan, Gerrits, Walsh, & Van Puijenbroek, 2005). The reporting of the event related to other drugs in the database is used as a proxy for the background occurrence of the event (Poluzzi, Raschi, Piccinni, & De Ponti, 2012). To put it in another way, different disproportionality measures quantify the extent to which a given event is frequently reported with a given drug by estimating the observed-to-expected co-occurrence ratio. Those drug-event pairs that are significantly different from background reporting occurrence may reflect credible signals
that need further investigation. However, reported ADRs do not necessarily accurately reflect the usage of drugs and the incidence of ADRs in the population (Poluzzi et al., 2012), the significant pairs may represent reporting tendencies (Manfred Hauben & Reich, 2005). Commonly used disproportionality measures include the proportional reporting ratio (PRR) (Evans, Waller, & Davis, 2001), the reporting odds ratio (ROR) (Puijenbroek et al., 2003), Bayesian confidence propagation neural network (BCPNN, information component (IC) is the statistical score) (A. Bate et al., 1998; Lindquist et al., 1999, 2000; Noren et al., 2006), multi-item gamma Poisson Shrinker (MGPS, empiric Bayes geometric mean (EBGM) is the statistical score) (W. DuMouchel & Pregibon, 2001; W DuMouchel, 1999), and others (Table 2-1). Different comparison references are used to estimate the expected ADR occurrences in frequency probability based measures. The observed for all drugs and the observed for all other drugs are used to estimate the expected occurrence of the ADR for PRR and ROR respectively (Poluzzi et al., 2012). With Bayesian methods, if a drug and an ADR is independent, then their joint probability to the product of the individual probabilities will equal to 1 (Poluzzi et al., 2012). This inference hypothesis is used to calculate and interoperate observed-to-expected co-occurrence ratio. Roux et al. (E. Roux et al., 2003) evaluated these statistical models on simulated datasets (constructed by SRS modeling for arbitrary selected 150 drugs and 100 side effects) by comparing the percentage of false positive signals among given drug-event combinations (Emmanuel Roux, Thiessard, Fourrier, Begaud, & Tubert-Bitter, 2005). The false positive rates varied from 1.1% to 53.4%; and EBGM, IC and Chi-square models seemed to have better performances (Emmanuel Roux et al., 2005). Puijenbroek et al. (van Puijenbroek et al., 2002) also evaluated different measures and found out that sensitivity was high with respect
to the reference measure when a combination of point- and precision estimates was used. For example, in the PRR measure, the PRR is the point estimate; and its lower limit of 95% Confidence Interval (CI) is the precision estimate.
Table 2-1: Quantitative data mining procedures used as disproportionality measures in SRS (Balakin & Ekins, 2009; M. Hauben & Bate, 2009; Manfred Hauben et al., 2005; Emmanuel Roux et al., 2005; van Puijenbroek et al., 2002)

<table>
<thead>
<tr>
<th>Measure of Association and Data source</th>
<th>Formula</th>
<th>Measure of importance</th>
<th>Probabilistic Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Frequentist Approaches</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportional Reporting Ratio (PRR)</td>
<td>$PRR = \frac{a}{(a + b)c/(c + d)}$</td>
<td>$PRR - 1.96 SE &gt; 1$ or $PRR \geq 2, a \geq 3, \chi^2 \geq 4$</td>
<td>$P(A</td>
</tr>
<tr>
<td>(Evans et al., 2001)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK Yellow Card database, Medicines Control Agency (MCA)</td>
<td>$SE(\ln PRR) = \sqrt{\left(\frac{1}{a} - \frac{1}{a + b} + \frac{1}{c} + \frac{1}{c + d}\right)}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$95% CI = e^{\ln (PRR) \pm 1.96 SE(\ln PRR)}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reporting Odds Ratio (ROR) (Puijenbroek et al., 2003)</td>
<td>$ROR = \frac{(a / c)}{(b / d)} = \frac{ad}{bc}$</td>
<td>$ROR - 1.96 SE &gt; 1$</td>
<td>$P(A</td>
</tr>
<tr>
<td>Netherlands Pharmacovigilance Centre Lareb</td>
<td>$SE(\ln ROR) = \sqrt{\left(\frac{1}{a} + \frac{1}{a + b} + \frac{1}{c + d}\right)}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$95% CI = e^{\ln (ROR) \pm 1.96 SE(\ln ROR)}$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 2-1: Continued

<table>
<thead>
<tr>
<th>Measure of Association and Data source</th>
<th>Formula</th>
<th>Measure of importance</th>
<th>Probabilistic Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bayesian Approaches</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relative Risk (RR)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$RR = \frac{n_{ij}}{E_{ij}} = \frac{a}{(a + b)(a + c)}$</td>
<td>$IC = \log_2 \left( \frac{a(a + b + c + d)}{(a + b)(a + c)} \right)$</td>
<td>$IC - 2SD &gt; 0$</td>
<td>$\log_2 \left( \frac{P(A,D)}{P(A)P(D)} \right)$</td>
</tr>
<tr>
<td>WHO BCPNN and FDA MGPS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$RR = \frac{a(a + b + c + d)}{(a + b)(a + c)}$</td>
<td>$EBGM \ 05 &gt; 2$</td>
<td>$\frac{P(A</td>
</tr>
</tbody>
</table>

A: ADR; D: Drug; BCPNN: Bayesian Confidence Propagation Neural Network; MGPS: Multi-item gamma Poisson Shrinker
2.2.2 Taxonomic reasoning

In addition to the use of quantitative techniques, ontology-based reasoning in combination with quantitative data mining methods has been shown to be able to improve signal detection from SRSs (J. S. Almenoff et al., 2007; Bousquet, Henegar, Louêt, Degoulet, & Jaulent, 2005; Bousquet, Trombert, Kumar, & Rodrigues, 2008; Henegar, Bousquet, Lillo-Le Louët, Degoulet, & Jaulent, 2006; Mera, Beach, Powell, & Pattishall, 2010; Nadkarni, 2010). Different methods of ontological reasoning were conceived. Two examples, coined terminological reasoning by subsumption and approximate matching (Bousquet et al., 2005), are summarized in the following.

In a terminological hierarchy, when multiple subclasses are connected to one superclasses, it is said that the superclass subsumes all of the subclasses. Reasoning by subsumption then means that a given class and all of its subsumed classes are considered to be one entity. The new entity then constitutes all of the signals of the individual classes into one, increasing its number of occurrence. In SRSs, ADRs are usually coded using the Medical Dictionary for Regulatory Activities (MedDRA) terminology, and MedDRA’s hierarchy was used to facilitate terminological reasoning by subsumption.

In approximate matching, a new concept is built as the disjunction of all sets of possible characteristics. For example, Hepatitis is defined as “((hyperbilirubinemia AND ALAT increased) OR (ASAT increased AND cholestasis) OR (jaundice AND cytolysis))” (Bousquet et al., 2005). This is another way to group cases to increase the number of occurrences of identifiable drug/ADR associations.

These and other approaches and combinations of both approaches have been shown to improve signal detection by exploiting ontological properties with statistical data mining.
methods (J. S. Almenoff et al., 2007; Bousquet et al., 2005, 2008; Mera et al., 2010; Nadkarni, 2010).

2.3 Predicting drug side effect relations

2.3.1 Machine learning algorithms

Other researchers have attempted to predict which drugs cause which side effects using machine learning approaches with structured knowledge concerning the drugs and ADRs as features. Liu et al. applied five supervised machine learning (ML) algorithms with drug features drawn from PubChem (chemical substructures), DrugBank (drug target, transporters, and enzymes), KEGG (pathway) and SIDER itself (drug indication and side effects). A ML classifier was built for each SIDER ADR. The classifiers were then evaluated on 832 SIDER drugs (for which DrugBank IDs could be found) using five-fold cross-validation. Of the algorithms evaluated, the support vector machine (SVM) algorithm performed best with an AUC of 0.9054 on the SIDER dataset as a whole (Liu, Wu, et al., 2012).

In other work (Huang, Wu, & Chen, 2011), the features were selected from drug-target information from DrugBank and drug non-target protein information from an expanded protein-protein interaction (PPI) network derived from gene ontology (GO) annotations. A SVM classifier was built for each ADR and cross-validated on a known side effect resource (SIDER, version 1).

In related work, decision tree models were derived for specific end organs using features derived from the chemical structure of the agents in the SIDER set. The accuracies of
decision trees for allergic, renal, central nervous system (CNS) and hepatic ADRs ranged from 78.9% to 90.2% (Hammann, Gutmann, Vogt, Helma, & Drewe, 2010).

In addition to supervised ML algorithms, unsupervised learning using topic models has also been used to analyze drug labels and group drugs by topics that are associated with the same side effects (Bisgin, Liu, Fang, Xu, & Tong, 2011).

### 2.3.2 Regression and correlation analysis

Pauwels and colleagues applied sparse canonical correlation analysis (SCCA) utilizing the chemical substructure of drugs to predict side effects (Pauwels, Stoven, & Yamanishi, 2011). Features representing drugs’ chemical substructures were extracted from PubChem. Side effects were extracted from SIDER. Canonical correlation analysis (CCA) is a way of measuring the linear relationship between two multidimensional variables and customized to analyze the relationship between the group of chemical substructure and the group of side effects and to find what chemical substructure can represent what side effect. Highly correlated sets of chemical substructures and side effects were retrieved, and predictive scores associating substructures to side effects were calculated. SCCA is a variant of canonical correlation analysis that aims to minimize the size of the set of features utilized so as to facilitate explanation. The AUC for SCCA is 0.8932 for predicting side effects.

In another study, logistic regression was used to predict ADRs using as features chemical substances from PubChem and BioAssay (a database of bioactivity screens of chemical substances). Performance was evaluated for side effects related to specific organ systems, and was best for immune system disorders (AUC=0.92) and blood and lymphatic system disorders (AUC=0.79) (Pouliot, Chiang, & Butte, 2011).
2.3.3 Mining literature to predict drug/ADR associations

Currently, systematic literature review is utilized as an in-depth investigation to provide scientific evidence to confirm a specific drug/ADR relationship or support regulatory decisions (for example, drug withdrawal) in the pharmacovigilance process (Arnaiz et al., 2001; Y. K. Loke, Kwok, & Singh, 2011; Yoon K Loke, Price, & Herxheimer, 2007; P. Lopes et al., 2013; Marie et al., 2008; Olivier & Montastruc, 2006). However, evidence in the biomedical literature (e.g. case reports, clinical trials) may appear before an ADR is recognized. This implies that analyzing the literature to predict drug/ADR associations may complement current drug safety methods (disproportionality measures of SRS). This aspect of pharmacovigilance has been explored in a few studies only (Alomar, Hourani, & Sulaiman, 2008; Deftereos, Andronis, Friedla, Persidis, & Persidis, 2011; Shetty & Dalal, 2011).

There are two major investigative perspectives. One perspective involves using the literature as a data source for ADR case reports, and then applying statistical models to mine the drug/ADR associations (Alomar et al., 2008; Shetty & Dalal, 2011). Shetty et al. (Shetty & Dalal, 2011) retrieved literature about side effects by searching PubMed using NLM’s Medical Subject Headings (MeSH) index, and then filtering irrelevant articles (for example articles discuss only treatments) using a Lasso-based document relevance classifier (Genkin, Lewis, & Madigan, 2007). With the contingency table built from drug/ADR occurrence across all relevant citations, a disproportionality analysis was applied to find statistically significant drug/ADR associations. This model discovered 54% of all detected FDA warnings from labeling and also confirmed the association between
rofecoxib and heart disease with literature before 2002 while rofecoxib was recalled in 2004.

The second perspective is based on the paradigm of literature-based discovery (LBD), and involves processing the literature to uncover indirect relationships between drugs and ADRs. Using LBD for pharmacovigilance has not been attempted until recently Hristovski et al. (Hristovski et al., 2006) model LBD discovery patterns to analyze semantic predications and find therapeutic relations for drugs and a disease. To date, LBD-inspired pharmacovigilance research has concerned providing justification for a connection between a drug and an ADR, or investigating possible mechanisms to explain an observed side effect using the literature (Ahlers, Hristovski, Kilicoglu, & Rindflesch, 2007; Hristovski, Burgun-Parentoine, Avillach, & Rindflesch, 2012a). However, this approach has not yet been evaluated as a means to predict side effects on a large scale. I will elaborate further on this aspect in Section 2.7.1.

2.3.4 Statistical data mining from EHR for signal detection

Using EHR data for pharmacovigilance is an active research area (Wang et al., 2009). Disproportionality measures applied to SRSs have been applied to EHR data also (Zorych, Madigan, Ryan, & Bate, 2011; Liu, McPeek Hinz, et al., 2012). However, these algorithms were designed for reporting systems to evaluate whether a given event is disproportionally “reported” with a given drug. As side effects may not be explicitly reported in the EHR, the applicability of disproportionality measures to EHR data must still be evaluated (Zorych et al., 2011). Zorych et al. (Zorych et al., 2011) hypothesize that the observed co-occurrence of a drug and an event can be regarded as though they were reported as a spontaneous reporting case, to produce drug-event contingency tables. Using this
approach, existing disproportionality algorithms were applied to both real world and simulated EHR data. Results show that the disproportionality algorithms can find legitimate drug/ADR associations; they also produce a large number of false positive associations. Liu et al. (Liu, McPeek Hinz, et al., 2012) assessed disproportionality measures on drug-laboratory test ADR associations derived from EHR data. This study further demonstrated that disproportionality measures can be applied to analyze EHR data for pharmacovigilance. Of note, these EHR data (Zorych et al., 2011; Liu, McPeek Hinz, et al., 2012) were structured.

However, most clinical information also includes unstructured narratives, which contain valuable patient data. Several researchers (Cao, Hripcsak, & Markatou, 2007; Cao, Markatou, Melton, Chiang, & Hripcsak, 2005; Wang et al., 2009) have investigated measuring co-occurrence within the narrative section of the EHR as a means to detect possible drug/ADR associations. This is accomplished by calculating above-chance occurrences of drug and problem pairs in the clinical notes, where each note represents a clinical encounter. Associations among entities are estimated using statistical methods (Chi-Square statistics) based on their co-occurrence statistics in the clinical notes. Co-occurrence statistics (Christopher D. Manning & Hinrich Schütze, 1999) have been successfully applied in the computational linguistics domain, for example, for the purpose of word sense disambiguation (Yarowsky, 1995), automatic generation of semantic lexicons (Roark & Charniak, 1998) and synonym mining (Turney, 2001). In addition to word-level co-occurrence statistics, concept-level co-occurrence has also been used to discover associations between concepts (Cao et al., 2005). If a drug causes an ADR, then this drug and this related problem are more likely to appear together than random
combinations of drugs and adverse events in clinical notes (Cao et al., 2005). It has been demonstrated that co-occurrence statistics can be derived from unstructured EHR data to detect possible side effects (Wang et al., 2009).

2.4 Pharmacovigilance challenges from my perspective

The above review of the literature highlights two major challenges: inefficient and incomplete SRS data and the lack of computer-assisted causality assessment in pharmacovigilance practice. As discussed in section 2.1.3, EHR may compensate for the limitations of SRSs and provide the opportunity for active surveillance. As argued in (Anderson & Borlak, 2011), there is a lack of causality assessment in pharmacovigilance practice. While expert clinical review is designed to verify potential ADRs, it is a human-intensive and time-consuming process. The available human resources are inadequate to review the large amount of noisy signals detected in SRS and EHR data, creating a bottleneck in the pharmacovigilance process. More research is needed to develop methods to automate, or assist with, the knowledge-intensive task of expert clinical review. In the section that follows, I will discuss emerging methodologies that I propose adapting in order to address these limitations: EHR data can be used as an alternative data source for pharmacovigilance and scalable LBD methods based on distributional semantics can be utilized to assist signal evaluation.

2.5 Signal detection from EHR

An EHR system presents the possibility of a real-time, continuous approach to drug surveillance (Brown et al., 2007; Trifirò et al., 2009; Wang et al., 2009). It has been
estimated that Vioxx would have been withdrawn in just 3 months instead of 5 years if the EHR data of 100 million patients had been available for drug monitoring (McClellan, 2007; Pray, Robinson, & Translation, 2007; Trifirò et al., 2009; Wagner et al., 2006). In addition to review the statistical mining from EHR (Section 2.3.4), I continue to review the recent proposed framework using EHR for pharmacovigilance from a pilot study (Wang et al., 2009).

Wang et al. (Wang et al., 2009) proposed a pharmacovigilance framework utilizing an EHR system. The basic procedure is: (1) Drug-relevant data (both structured and unstructured) are extracted from the EHR. A Natural Language Processing (NLP) tool is used to parse free-text clinical notes because unstructured data may include richer patient information; (2) The drug relevant data are coded and mapped to a standard terminology for further analysis; (3) Drug and related event pairs are selected by filtering indication relationships and excluding incorrect temporal order of events (i.e. side effect before drug administration); (4) Selected drug-event pairs are analyzed using statistical data mining algorithm (Chi-square statistics) to obtain a ranked ADR list for all respective drugs based on the strength of the statistics; (5) For each drug’s ranked ADR list, an empirical threshold of significance level is adjusted by volume tests using the $P$-value plot (Schweder & Spjøtvoll, 1982; Cao et al., 2005, 2007; Wang et al., 2009) to determine possible signals for each drug.

Essentially, there are two major tasks required to use EHR data for pharmacovigilance. The first task is to process and clean EHR data to get drug-event combinations and retrieve these combinations’ co-occurrence data using informatics tools. Second is to conduct
statistical analysis to analyze these co-occurrence data of drugs and events to detect the significant drug/ADR pairs.

ADR signals have relatively low incidence in comparison with background information in the EHR (for example, therapeutic relationships are more common) because EHRs are not designed for ADR monitoring. This leads to noise, which means the application of basic statistical methods alone will detect many false signals. Thus estimating a threshold to find stronger associations maybe difficult (Cao et al., 2007; Wang et al., 2009). However, pharmacovigilance from EHR can provide good signal candidates for further investigation.

2.6 Expert clinical review and causality assessment in pharmacovigilance

2.6.1 Expert clinical review in pharmacovigilance

Since the beginning of pharmacovigilance practice (mid 1960s), expert clinical review has been viewed as essential to assess causality during the signal evaluation process (Agbabiaka, Savović, & Ernst, 2008; Andrews & Moore, 2014; Andrew Bate, 2003). Detected signals are reviewed by one or more experts who are often specialists, with expertise related to the ADR concerned (Andrews & Moore, 2014). This is referred to as expert judgement or global introspection -- an expert expresses a judgement about possible drug causation after having taken into account all the available and relevant information on the considered case (Arimone et al., 2007). Expert judgment assessments are generally expressed in terms of a qualitative probability scale (Andrews & Moore, 2014). For example, the WHO causality categories are ‘Certain’, ‘Probable’, ‘Possible’, ‘Unlikely’, ‘Conditional/unclassified’, and ‘Unassessable/unclassifiable’(Meyboom, Hekster, Egberts, Gribnau, & Edwards, 1997). During the assessment, specific structured
approaches have been used to guide the experts from which perspective to evaluate the possible ADR. For instance, the Swedish regulatory agency (Wiholm, 1984) recommends the experts to assess possible ADRs from the following aspects: (1) the temporal sequence, (2) previous information on the drug, (3) dose relationship, (4) response pattern to drug, (5) rechallenge, (6) alternative aetiological candidates, and (7) concomitant drugs.

2.6.2 Assessment of causality

To address the assessment of causality, general principles exist that can be applied to evaluate the causality of potential ADRs (Shakir & Layton, 2002). The theoretical basis for these principles was proposed by Sir Austin Bradford-Hill in 1965 (Hill, 1965). Bradford-Hill, an English epidemiologist and statistician, was the first to demonstrate that cigarette smoking contributes toward lung cancer using what are now referred to as the “Bradford-Hill criteria” (Richard Doll, 1994). The Bradford-Hill criteria provide viewpoints from which to evaluate evidence indicative of causality and have been widely used in pharmacoepidemiology. These criteria are named ‘strength’, ‘consistency’, ‘specificity’, ‘temporality’, ‘biological gradient’ (referring to dose-response relationships), ‘plausibility’, ‘coherence’, ‘experimental evidence’, and ‘analogy’ (Hill, 1965; Kleinberg & Hripcsak, 2011; Ward, 2009). Since then, the criteria have been widely used in epidemiology and may be applied to assess the causality of drug/ADR relationships (Anderson & Borlak, 2011; Perrio, Voss, & Shakir, 2007; Shakir & Layton, 2002). Three of these criteria seem particularly pertinent to the development of pharmacovigilance methods in our study:

- The strength criterion reflects that strong associations are more likely to be causal than weak associations (Shakir & Layton, 2002). Quantitative statistical data
mining methods evaluate ADR signals from the strength of association point of view.

- The *plausibility* criterion relates to evidence about mechanisms that may be involved to support a causal relationship.
- The *coherence* criterion relates to the consistency of the hypothesis in question with contemporary medical knowledge.

2.7 LBD and literature NLP annotation tools

2.7.1 LBD and discovery patterns

Processing published biomedical literature to uncover implicit relationships among entities is referred to as LBD (Bruza & Weeber, 2008; Swanson, 1986b, 1987; Weeber, Klein, de Jong-van den Berg, & Vos, 2001). LBD involves finding new knowledge by analyzing the literature, rather than through scientific experimentation. This is accomplished by identifying hidden connections between entities described in the published literature (Swanson, 1986a, 1986b). The origins of LBD may be traced to the serendipitous discovery that fish oils can be therapeutically useful in the treatment of Raynaud’s syndrome (poor circulation in the peripheries) by information scientist Don Swanson (Swanson, 1986a, 1986b). Weeber et al. described the two types of LBD (Weeber et al., 2001).

One type, referred to as “open LBD”, starts from a known term or concept (generally called A, although also referred to as C in Swanson’s early work) and tries to find an interesting hypothesis in the form of a previously unrecognized connection to some other term. If an article argues that A is associated with B and a second article mentions that B is associated with C, A may treat C. For example dietary fish oil (A) affects platelet aggregation, blood
viscosity and vascular reactivity (B), and these biological factors (B) play a role in Raynaud's syndrome (C) (Swanson, 1986a). Consequently, it is reasonable to hypothesize that A treats C. The open LBD process proceeds from the source term A to an unknown target term C and culminates in the generation of a new hypothesis.

The second type of LBD is referred to as “closed LBD”. In a closed LBD process the goal is to evaluate an existing hypothesis. Closed LBD starts with known terms A and C, with the goal to identify intermediate terms B that provide the bridge between A and C (Weeber et al., 2001). For example, in 1988 Swanson found intermediate concepts to explain a hypothetical relationship between migraine and magnesium (Swanson, 1988). Smalheiser and Swanson used closed LBD to propose an explanation for epidemiologic evidence that estrogen might protect against Alzheimer’s disease (Smalheiser & Swanson, 1996).

LBD methodologies generally utilize statistical information derived from the frequency with which terms, or discrete concepts extracted from the literature using automated tools (e.g. MetaMap) or assigned to it by human annotators (Srinivasan & Rindflesch, 2002), co-occur (Hristovski et al., 2006; Yetisgen-Yildiz & Pratt, 2006). This has been referred to as the co-occurrence model (Sehgal, Qiu, & Srinivasan, 2008). These co-occurrence statistics are interpreted by correlation-mining and ranking algorithms (Yetisgen-Yildiz & Pratt, 2006, 2009).

A limitation of these methods is that they generally do not consider the nature of the relationship between the terms or concepts concerned. To address this limitation, Hristovski and his collaborators (Hristovski et al., 2006) propose using semantic relations to eliminate spurious relationships introduced by frequently co-occurring concepts that are not meaningfully related. In their initial work, the semantic relations concerned were
extracted from the literature by two NLP systems: SemRep (Rindflesch & Fiszman, 2003) and (specifically to extract phenotypic information) BioMedLEE (L. Chen & Friedman, 2004). Their approach involved the specification of “discovery patterns”, patterns of relationships between concepts that may indicate an implicit therapeutic relationship (Hristovski, Friedman, Rindflesch, & Peterlin, 2008). These conditions can be specified as sets of semantic predicates. For example, Ahlers et al. (Ahlers, Hristovski, et al., 2007) defined the May_Disrupt pattern as follows:

Substance X <inhibits> Substance Y

Substance Y <causes|predisposes|associated with> Pathology Z

Substance X <may disrupt> Pathology Z

Variants of this approach have been applied to generate or support the hypotheses that fish oil treats Raynaud’s disease (Hristovski et al., 2006), insulin treats Huntington disease (Hristovski et al., 2006), and antipsychotic agents prevent cancer (Ahlers, Hristovski, et al., 2007). Recently, this approach was also adapted to provide evidence to support the plausibility of an observed drug/ADR association (Hristovski et al., 2012a; Hristovski, Burgun-Parenthoine, Avillach, & Rindflesch, 2012b), providing proof-of-concept that LBD methods can be applied within the problem domain of pharmacovigilance. Regardless of the application domain, knowledge used to populate discovery patterns is extracted from the biomedical literature using NLP.

2.7.2 Literature NLP annotation tools

MetaMap and SemRep are two examples of tools that have been used to extract information encoded in the biomedical literature. MetaMap is a widely-used NLP tool that identifies concepts from the Unified Medical Language System (UMLS) in biomedical text (A. R.
Aronson & Lang, 2010; A. R. Aronson, 2001). SemRep (Rindflesch, Fiszman, & Libbus, 2005; Rindflesch & Fiszman, 2003) is a rule-based NLP tool (A. R. Aronson & Rindflesch, 1998) that draws on concepts extracted by MetaMap and medical domain knowledge in the UMLS to extract semantic predications (Rindflesch & Fiszman, 2003). Its input consists of sentences from the literature; its output is a series of semantic predications identified in the respective text. A semantic predication is a subject-predicate-object triple sentence in which the subject and object are UMLS concepts and the predicate is a semantic relationship. For example, metformin (UMLS Concept C0025598) TREATS diabetes mellitus (C0011849) is a semantic predication extracted from the phrase “Treatment of diabetes mellitus with metformin”. Evaluations of SemRep reveal a precision between 0.73 and 0.81, and a recall of 0.55 on the biomedical literature (A. R. Aronson & Rindflesch, 1998; Kilicoglu, Fiszman, Rosemblat, Marimpietri, & Rindflesch, 2010; Rindflesch & Aronson, 2002). Semantic predications benefit the LBD process in several respects. The additional information provided by semantic predications makes the LBD results easier to interpret. In addition, it has been noted that a large number of uninformative co-occurrences must be manually reviewed when LBD is based on lexical statistics alone (Lindsay & Gordon, 1999). In contrast, semantic predications provide the means to isolate relationships between concepts that are logically connected in a meaningful way.

2.8 Semantic Vectors (Cohen & Widdows, 2009)

Regardless of whether co-occurrence relations or discovery patterns are used, LBD systems must explore large numbers of co-occurring terms or possible reasoning pathways to identify explanatory hypotheses (for closed discovery) or previously unrecognized
relationships (for open discovery). Consequently, the process of LBD can be computationally expensive, and thus faces scalability issues in the context of the rapid growth of the biomedical literature. In contrast, the field of distributional semantics has produced corpus-derived statistical models that can measure the relatedness between two concepts by comparing vector representations of these concepts, called semantic vectors, that are derived from the contexts they have occurred in (Cohen & Widdows, 2009), without the need to explicitly explore co-occurring concepts once the initial model has been generated. Consequently, several authors have explored the use of distributional models for LBD (Cohen et al., 2010; Cole & Bruza, 2005; Gordon & Dumais, 1998).

2.8.1 Overview of semantic vectors

These geometrically motivated models of distributional semantics represent terms or concepts as high-dimensional vectors derived from the contexts in which they have occurred. Relatedness between a pair of terms or concepts is then estimated from their representing vectors’ similarity (Cohen et al., 2010). The vector components can be binary, ternary, real, or complex values (Kanerva, 2009; Widdows & Cohen, 2012).

The overall methodology of semantic vectors is to build vector representations for terms and documents. Different methods exist to compute these vectors. For example, document vectors can be built from term-by-context matrices (Cohen & Widdows, 2009); in Latent Semantic Analysis (LSA) the initial representation is a term-document matrix derived from a corpus (Landauer & Dumais, 1997).

Overall, document vectors are derived from the vector representations of the terms that they contain and as a consequence, documents with similar terms have similar vector representations. A mathematically computed distance between vectors is then
representative of the similarity of the documents they represent. A key aspect is that the abstract concept of meaning or semantics of a concept is converted into a metric that can be computationally exploited.

Term-by-context matrices or term-document matrices can be large for text corpora with a high number of terms and documents. PubMed/MEDLINE for example has both dozens of millions of terms and documents. Deriving vector representations with term-by-context or term-document matrices can thus become computationally unfeasible. Methods to circumvent this problem by reducing the vector dimensionality have been conceived and shown to effectively preserve the meaning of the represented concepts as described in the next section.

2.8.2 Scalability and random indexing (RI)

RI, a relatively recent development for dimension reduction, supports the derivation of semantic distance from large corpora at minimal computational expense (Cohen & Widdows, 2009). RI further improves the scalability of distributional methods by avoiding computationally intensive approaches to dimension reduction of the original term-by-context matrix (Kanerva, Kristofersson, & Holst, 2000; Kanerva, 2010). The algorithm’s computational complexity scales linearly with increasing size of the input data. It can be incrementally updated as new documents are added without retraining the whole dataset; thus it is applicable to large corpora such as MEDLINE. Two variants of RI have been applied to LBD in our previous work, which can efficiently infer and identify therapeutic relationships. One is Reflective Random Indexing (RRI) (Cohen et al., 2010), which models co-occurrence. The other one is Predication-based Semantic Indexing (PSI) (Cohen, Widdows, Schvaneveldt, Davies, et al., 2012), which implements discovery
patterns in vector space. On account of their scalability, these models permit inference on a scale that would be prohibitively time-consuming if explicit exploration of all possible reasoning pathways were attempted. This is accomplished through a mechanism known as “indirect inference” (Landauer & Dumais, 1997), which enables distributional models to find meaningful connections between terms that do not co-occur with one another directly, without the need to explore intervening terms explicitly. Further details about these two LBD distributional semantics models (RRI and PSI) are provided in Section 4.2.1 and 4.2.2 respectively.

2.9 Research opportunities

2.9.1 Active drug surveillance with EHR

With the broader availability of EHR data and the building of data repositories by integrating different EHR systems (FDA Science Board Subcommittee, 2011; P. Lopes et al., 2013; Platt et al., 2009; Stang et al., 2010), an active and real time pharmacovigilance surveillance may be achieved in the near future. This poses informatics challenges from the data integration perspective. With respect to EHR signal detection, the challenge of improving signal detection is likely to be a research focus. This can be achieved using statistical data mining methods, and by improving the accuracy of true drug/ADR co-occurrence data using informatics methods.

2.9.2 Using distributional semantic models to find plausible drug/ADR associations

Signal evaluation is still a key pharmacovigilance challenge (P. Lopes et al., 2013). The proposal of substantiating and verifying ADR signals by analyzing the literature and existing drug-related knowledge bases in an automated fashion raises many research
questions. With the scalability of newer distributional semantics methods, and their capability of modeling the indirect relationships between entities using literature, I posit that LBD distributional semantics can retrieve evidence from the literature to establish the plausibility of connections between drugs and ADRs. This will assist the pharmacovigilance evaluation process by providing relevant evidence for domain experts to consider for causality assessment in signal evaluation.

In the chapters that follow, I will evaluate this hypothesis by using state-of-the-art statistical algorithms to analyze an EHR system and identify statistically significant drug/ADR associations. Scalable LBD models based on distributional semantics are designed and built to leverage knowledge from the biomedical literature to identify plausible drug/ADR associations. An evaluation is conducted to determine the validity of each developed method.
Chapter 3: Signal Detection from Outpatient Electronic Health Record Data Using Statistical Mining

Research suggests data collected by SRS are limited by long time latency, incorrect or incomplete clinical information, underreporting and reporting bias (Alvarez-Requejo et al., 1998; Hasford et al., 2002). Clinicians and researchers have also utilized existing healthcare data sources such as Electronic Health Records (EHR) to attempt to identify previously unreported adverse drug reactions (ADRs) (Haerian et al., 2012; Harpaz et al., 2013; Trifirò et al., 2009; Wang et al., 2009). However, these data are inherently noisy as drugs and potential side effects may co-occur in the EHR for many reasons. In addition, the EHR often contains free-text data, and the accuracy of Natural Language Processing (NLP) tools is not perfect. New methods are required to selectively identify potentially hazardous drug/ADR associations. Consequently, the development of computational approaches to more accurately detect potential side effects is currently an active area of research (Bauer-Mehren et al., 2012; Deftereos et al., 2011; Friedman, 2009; Oliveira et al., 2013; Shetty & Dalal, 2011). These approaches have predominantly focused on improving signal detection using statistical methods, machine learning or some combination thereof.

Statistical methods employed mostly involve disproportionality analysis. Other statistical algorithm based on co-occurrence has also been explored to detect possible side effects from unstructured clinical notes. Both statistical methods are based on an independence
model (Wang et al., 2009; Zorych et al., 2011). Disproportionality measure was developed from SRS data and has been tested to be able to retrieve drug/ADR associations from EHR data (Liu, McPeek Hinz, et al., 2012; Zorych et al., 2011) under the premise that the observed co-occurrence of a drug and an event can be considered as the reported a possible drug/ADR association (Zorych et al., 2011). Other statistical algorithm based on co-occurrence, originated from detecting the above-chance frequent occurrence of two entities from a text corpus, has also been demonstrated to be effective in retrieving possible drug/ADR associations from clinical corpus (discharge summaries in (Wang et al., 2009)). The motivation of using this algorithm is based on the hypothesis that a drug entity and a possible problem entity are more likely to appear together than random combinations of any drug entity and any problem entity (Cao et al., 2007; Wang et al., 2009).

However, EHR data was not captured for the purpose of reporting side effects, and consequently there exists information other than drug related problems, e.g. drug treatment information (symptoms, indication, etc.). So researchers need to select a related cohort for investigating ADRs using EHR data. Liu et al. (Liu, McPeek Hinz, et al., 2012) selected drugs and corresponding abnormal laboratory results as possible side effects from inpatients structured clinical data as the cohort. Wang et al. (Wang et al., 2009) utilized a NLP tool to process unstructured clinical notes and only selected drug and possible side effects related to the processed sections as the cohort. Consequently, statistical methods and machine learning tools are utilized together to analyze unstructured clinical notes for finding possible drug/ADR associations.

Structured (Liu, McPeek Hinz, et al., 2012) and unstructured (Wang et al., 2009) inpatient EHR data and structured outpatient (Honigman, Light, Pulling, & Bates, 2001; Honigman,
Lee, et al., 2001) EHR data have been demonstrated to be a possible data source to identify ADRs. However, to our knowledge, unstructured outpatient EHR data hasn’t been used for ADRs identification. OMOP preliminary results suggest that the performance of using EHR systems varies by data source (FDA Science Board Subcommittee, 2011). In this study, an unstructured outpatient EHR data is evaluated to be a feasible data source for detecting ADRs. Both disproportionality measures and other statistical algorithm based on co-occurrence are used to analyze the outpatient EHR data and find possible drug/ADR associations. Corresponding workflow is discussed.

3.1 Materials

3.1.1 Clinical data warehouse (CDW)

The CDW was developed by the Center for Clinical and Translational Science (CCTS) at the University of Texas Health Science Center at Houston. The outpatient EHR system for UT Physicians (Allscripts), is hosted on CDW for clinical research usage. It contains medical records on approximately 364,000 patients treated by UT Physicians (Saitwal et al., 2012). The batch used in this experiment is 20130130 containing about ten-year clinical data until January, 2013 and 2,603,279 outpatient clinical notes. This large sample size provided by the EHR system may provide enough power to detect small frequency ADRs. This work is qualified for an IRB exemption and has been approved by the Committee for the Protection of Human Subjects. This project refers to HSC-SBMI-12-0226 – “Detection of adverse drug events from electronic health records”.
3.1.2 Medical language extraction and encoding system (MedLEE)

MedLEE is a NLP system, designed to parse narrative patient reports into structured representations including UMLS codes for identified concepts (Friedman, Shagina, Lussier, & Hripcsak, 2004). Friedman et al (Friedman et al., 2004) report the recall of MedLee for UMLS coding of all terms from 150 randomly selected sentences -- selected from a text collection consisted of 818,000 sentences that were retrieved from de-identified patients’ discharge summaries -- as 0.77 compared with a reference standard which was determined by seven experts’ manual review. The precision of the system was reported as 89% for encoding a second set of 150 randomly selected sentences from the text collection then subsequently validated by experts. Since the overall evaluation was conducted on the records from the same institution in which the NLP system was developed, performance may be lower elsewhere as NLP systems often must be adapted to new contexts.

3.1.3 Side effect resource 2 (SIDER2 (Kuhn, Campillos, Letunic, Jensen, & Bork, 2010))

SIDER2 is a publicly available database containing information on marketed medicines and their known adverse reactions (Kuhn et al., 2010). SIDER2 was used to construct a dataset for our experiment and as a reference standard to confirm whether a predicted side effect is a true adverse reaction. SIDER2 contains 996 drugs, 4,192 side effects, and 99,423 drug/ADR pairs. Only those side effects and drugs that were contained in the EHR data were retained, leaving a total of 833 drugs, 2123 side effects, and 71,753 drug/ADR pairs.

3.1.4 Hierarchical relationships between concepts from UMLS knowledge base

The number of clinical notes for a problem was extended by grouping related concepts based on the UMLS semantic network and UMLS Metathesaurus, specifically the
MRREL.RRF file (U. S. National Library of Medicine, 2009). This file includes relationships between UMLS concepts found in the UMLS Metathesaurus, while the relationship can be synonym (SY), child (CHD), sibling (SIB), parent (PAR) and etc. By utilizing these ancestor-offspring hierarchical relationships (Figure 3-1), all offspring nodes for a SIDER2 ADR can be retrieved and subsequently are considered as related concepts, and their corresponding clinical notes are counted for the SIDER2 ADR. Take Breast Carcinoma as an example, Breast Carcinoma (C0678222) is a parent node of Mucinous Carcinoma of Breast (C1334807) and Lobular Carcinoma (C0206692), and a child node of Carcinoma (C0007097). This hierarchical relatedness can be used to group related concepts for Breast Carcinoma and consequently to count its clinical notes.

![Figure 3-1: The example ancestor-offspring hierarchy relationships between UMLS concepts](image)

3.1.5 MEDication-Indication (MEDI)

It was anticipated that co-occurring drugs and problems in the clinical notes might often be related therapeutically. Consequently, I evaluated the utility of an indication knowledge resource, MEDI (Wei et al., 2013), as a means to eliminate drug indications from consideration as potential side effects. MEDI is a medication indication resource that was extracted from a set of commonly used medication resources, including RxNorm,
MedlinePlus, SIDER2, and Wikipedia (Wei et al., 2013). MEDI drugs are represented by RxNorm codes, and indications are represented by ICD-9 codes. MEDI contains 3112 medications and 63,343 medication-indication pairs (MEDI-Complete). Additionally, the MEDI high-precision subset (MEDI-HPS) was also created by only including indications that are retrieved from RxNorm or at least two of the three other resources. MEDI-HPS contains 2136 medications and 13,304 medication-indication pairs. The estimated precision of MEDI-HPS is about 92%.

3.2 Methods

3.2.1 Annotating the EHR data using MedLEE

Drugs and co-occurring problems from medical records in the CDW were annotated by MedLEE. The medical records I utilized are free-text, unstructured clinical encounter documentations in outpatient setting. Each clinical summary was processed by MedLEE, resulting in an output that included UMLS-codes and semantic types for each extracted concept. Concepts denoted as a “medication” or “problem” were utilized as the source of potential side effects. Their occurrence at document level was recorded, to enable evaluation of the frequency with which a drug or a problem occurred in the EHR data.

3.2.2 Construct test data set

All SIDER2 drugs and side effects that are contained in the EHR data were utilized for the experiment. All selected drugs and ADRs are paired up to construct a SIDER2 drug/ADR test set. This resulted in a set of 1,768,459 drug/ADR relationships, 71,753 of which were true and 1,696,706 of which were false.
3.2.3 Retrieve co-occurrence data for SIDER2 test set

Co-occurrence data for the SIDER2 drug/ADR test set were retrieved from the EHR system and is referred to as original co-occurrence data. By using the ancestor-offspring hierarchical relationships in UMLS knowledge base, for each SIDER2 ADR, the EHR clinical documents that contain this concept or this concept’s all related offspring concepts are considered as ADR positive reports. This may also improve statistical power since the prevalence of some side effects is relatively low. By doing so, a second set of co-occurrence data were retrieved for each SIDER2 drug/ADR test set and is referred to as descendent co-occurrence data.

3.2.4 Statistical algorithms

3.2.4.1 Disproportionality analysis

After retrieving co-occurrence data for the drugs and ADRs that appear in the test set, a contingency table was constructed for each drug-problem pair. Disproportionality measures – reporting odds ratio (ROR), proportional reporting ratio (PRR), Yule’s Q (Yule), information component (IC), Multi-item gamma Poisson Shrinker (MGPS) -- were applied to identify statistically significant drug/ADR pairs (Table 2-1 in Chapter 2). If the observed occurrence / expected occurrence is more than a quantitative threshold (which varies across different disproportionality measures), the drug/ADR combination is considered as a signal.

3.2.4.2 Other statistical algorithm based on co-occurrence

To test independence, Chi-squared test can be applied to drugs and problem-related entities extracted from the EHR data using NLP (or structured data if available) to test the strength of the relationship between drugs and problems based on their paired and independent
observations in the dataset. There may exist interrelations between different drugs and different problems.

The Chi-squared value reflects the magnitude of the dependence. However, traditional Chi-squared statistics rejects most null hypotheses due to the large sample size (Cao et al., 2007) which results in false positives. Zorych et al. (Zorych et al., 2011) has shown that there are large amount of false positive signals detected from EHR data. In Chi-squared test, each drug/ADR pair is tested separately without taking into account the multiple comparison. Multiple comparison refers that multiple drug/ADR associations hypotheses are tested from the same dataset simultaneously (Ahmed, Dalmasso, et al., 2010). Researchers have proposed statistical approaches to control error rates in multiple hypothesis testing. Among these, the false discovery rate (FDR) was introduced to measure the multiple-hypothesis testing error and has been successfully used in large-scale genomic studies to control false positive results (Benjamini & Hochberg, 1995; J. D. Storey & Tibshirani, 2003; J. D. Storey, 2002). FDR avoids the over-conservativeness of the standard Bonferroni approach to multiple hypothesis testing, and has proven to be powerful in identifying sparse signals from a large number of tests (Cai & Sun, 2009). For example, FDR is used to detect differential gene expression in replicated DNA microarray experiments, where unknown dependencies are likely to occur (J. Storey & Tibshirani, 2001). On account of the proven utility of the FDR approach as a means to identify sparse signals from large datasets (Ahmed, Thiessard, Miremont-Salamé, Bégaud, & Tubert-Bitter, 2010), I also used the Chi-squared test augmented with the FDR as measuring significance threshold. The FDR approach (J. D. Storey & Tibshirani, 2003) determines the statistical significance of a q value for each tested pair. q values are estimated from p
values, which measure significance in terms of the false positive rate. The false positive rate is the rate at which null hypotheses are rejected (false drug/ADR associations are predicted as positive by a statistical model). For example, a false positive rate of 5% means that on average 5% of the false drug/ADR associations will be predicted as positive by the statistical model. The q value measures significance in terms of false discovery rate, the rate at which statistically significant signals are, in fact, false drug/ADR associations. Table 3-1 describes how overall accuracy or error is measured in m drug/ADR pairs testing.

Table 3-1: Overall error measure in FDR approach (J. D. Storey & Tibshirani, 2003)

<table>
<thead>
<tr>
<th>Tested significant (Positive)</th>
<th>Tested not significant (Negative)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Null hypothesis is True</strong></td>
<td>False Positive (FP)</td>
<td>True Negative (TN)</td>
</tr>
<tr>
<td>(no relationship between drug</td>
<td>α (Type I error)</td>
<td></td>
</tr>
<tr>
<td>and ADR)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Alternative hypothesis is</strong></td>
<td>True Positive (TP)</td>
<td>False Negative (FN)</td>
</tr>
<tr>
<td><strong>True</strong></td>
<td></td>
<td>β (Type II error)</td>
</tr>
<tr>
<td>(has relationship between drug</td>
<td></td>
<td></td>
</tr>
<tr>
<td>and ADR)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>S = 1 (FP+TP)</td>
<td>S = 0</td>
</tr>
<tr>
<td></td>
<td>(# of pairs tested significant)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{false positive rate} = E\left[\frac{F}{m_0}\right] \leq 0.05
\]

\[
\text{false discovery rate (FDR)} = E\left[\frac{F}{S}\right] = E\left[\frac{F}{F+T}\right] \leq 0.05
\]

\[
\text{Specificity} = \frac{m_0 - F}{m_0}
\]

\[
\text{Sensitivity} = \frac{T}{m_1}
\]

\[
\text{false discovery rate (FDR)} = E\left[\frac{m_0 \cdot (1 - \text{specificity})}{m_0 \cdot (1 - \text{specificity}) + m_1 \cdot \text{sensitivity}}\right]
\]
3.2.5 Filtering indications

To evaluate the utility of existing knowledge of indication, the two MEDI lists were applied to exclude indication relationships from the false pairs and tested using the top 5 best performing statistical models using the original SIDER2 test set.

3.2.6 Performance evaluation

The performance for different statistical algorithms were calculated and compared using the original SIDRE2 test set. The best five performing models were advanced for testing with different MEDI interventions and procedures to group related concepts.

3.3 Experimental design

The CDW EHR clinical narratives were processed and annotated by MedLEE (Friedman et al., 2004) with default setting. MedLEE extracts biomedical concepts from the EHR and categorizes these into semantic types based on large lexicon. Within MedLEE’s annotated output, those entities labels as “medication” and “problem” were considered as candidate drugs and ADRs respectively. SIDER2 (Kuhn et al., 2013) was used to construct a drug/ADR reference standard (SIDER2 drug/ADR test set) by pairing up all drugs and ADRs that exist in both the EHR data and SIDER2 to generate a set of 71,753 true and 1,696,706 false drug/ADR relationships. Since SIDER2 was extracted from package inserts by text mining tools, there may exist false drug/ADR associations because of the text mining error or over reporting of side effects (Duke J, Friedlin J, & Ryan P, 2011). Existing disproportionality measures and Chi-squared test with FDR for detecting drug/ADR associations from outpatient clinical notes were calculated and compared. In
this experiment, different thresholds (4, 6, 10, 25) of co-occurred drug/ADR reports were considered as the second condition to predict true drug/ADR associations.

Clinical narratives may have different concepts representing a SIDER2 ADR. Clinical notes containing these related concepts can be retrieved to expand the examples available for each SIDER2 ADR. In addition, a co-occurring drugs and problems may indicate therapeutic relationships, the MEDI lists were used to filter drug indication relationships from drug/problem candidates. Overall, my experiments concerned three methodological variants: (1) the choice of statistical measure; (2) increasing the examples available for each potential ADR by grouping related concepts for a SIDER2 problem; and (3) filtering known indication relationships.

3.4 Results

3.4.1 Experiment dataset

There are 2,603,279 clinical narratives (narrative notes in CDW updates to 01/30/2013) that were annotated by MedLEE. 2,325,614 notes that contain at least medication or problem were used for the experiment. There are 229 narrative types in this batch. 23.5% of the notes are labeled as “chart note” (Table 3-2). A chart note, also called as a progress note, is dedicated when an established patient is seen for a repeat visit.

In the EHR system, there are 7780 drugs and 10,670 problems. For SIDER2 set, there are 833 SIDER2 drugs that are also contained in the CDW, and 2123 SIDER2 side effects that are contained in the CDW. These overlapping drugs and problems between SIDER2 and CDW were used to build an evaluation set for the experiment. In this evaluation set, all
drugs were paired up with all problems. The true drug/ADR pairs are the pairs that are exist in SIDER2. The false drug/ADR pairs are the pairs that do not exist in SIDER2.

<table>
<thead>
<tr>
<th>clinical documentation type</th>
<th>Number of narrative notes</th>
<th>Percentage of total clinical notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chart Note</td>
<td>546,025</td>
<td>0.235</td>
</tr>
<tr>
<td>Clinical Note</td>
<td>180,192</td>
<td>0.077</td>
</tr>
<tr>
<td>Clinical Summary-RTF</td>
<td>144,486</td>
<td>0.062</td>
</tr>
<tr>
<td>Pre-Live Dictation.</td>
<td>120,295</td>
<td>0.052</td>
</tr>
<tr>
<td>Telephone Note</td>
<td>118,701</td>
<td>0.051</td>
</tr>
<tr>
<td>Nurse Note</td>
<td>113,253</td>
<td>0.049</td>
</tr>
<tr>
<td>Est Patient</td>
<td>96,567</td>
<td>0.041</td>
</tr>
<tr>
<td>Established</td>
<td>81,496</td>
<td>0.035</td>
</tr>
<tr>
<td>New Patient</td>
<td>63,110</td>
<td>0.027</td>
</tr>
<tr>
<td>UT Imaging.</td>
<td>60,419</td>
<td>0.026</td>
</tr>
</tbody>
</table>

3.4.2 Query the co-occurrence data for experiment datasets

For the drug/ADR pair in the evaluation set, the number of notes that contain the drug-of-interest and ADR-of-interest were counted to quantify co-occurrence. As discussed previously, to improve the statistical power, the notes that contain related problems-of-interest were also considered relevant notes when estimating the drug-problem co-occurrence counts for this problem (so any drugs co-occurring with a taxonomically related problem would be counted as though they had co-occurred with the index problem). On account of the two options for measuring associations with respect to related problems, there are two contingency tables for each experiment set (original vs. descendent co-occurrence).
3.4.3 Statistical results

3.4.3.1 Performance results

3.4.3.1.1 Performance of statistical algorithms with original co-occurrence data

The performance of predicting true positives, precision, recall and F-measure for different statistical algorithms with different co-occurrence threshold are compared and plotted in Figure 3-2. The average precision is 0.1315±0.0105. From this, I can anticipate that 10-15% of recovered signals from outpatient clinical notes can be true side effects. Chi-squared statistics with FDR threshold is the besting performing model with an F-measure of 0.1826. The best 5 performing statistical models based on F-measure are used in the following analysis.
Figure 3-2: Comparison of ability of different performance metrics (R-ROR, P-PRR, Y-Yule, C-Chi-square, F-Chi-square with FDR, M-MGPS) to recover SIDER2 side effects for different statistical algorithms with different drug/ADR co-occurrence threshold; top left, top right, bottom left, and bottom right are the comparison of number of true positives, precision, recall, and F-measure, respectively.

3.4.3.1.2 Effects of potential ontology intervention

The best 5 performing statistical models from previous results were used in comparing their performance between original and descendant co-occurrence data (Figure 3-3). With the descendant co-occurrence data, the algorithms detect more true positives, but this was counter-effect by more false positives retrieved and leads to the decreased precision. Overall, the F-measure decreases with the descendant co-occurrence data (Figure 3-4).
Figure 3-3: Comparison of the predicted drug/ADR association count comparing between original and descendant co-occurrence set for best five performing statistical models (F2-Chi-square with FDR and 25 co-occurrence; C-Chi-square with 25 co-occurrence; F1-Chi-square with FDR and 10 co-occurrence; P-PRR with 10 co-occurrence; Y-Yule with 10 co-occurrence); top left, top right, bottom left, and bottom right are the comparison of number of true positives, false positives, true negatives, and false negatives, respectively.
Figure 3-4: Comparison of ability of different performance metrics to recover SIDER2 side effects for best five performing statistical algorithms (F2-Chi-square with FDR and 25 co-occurrence; C-Chi-square with 25 co-occurrence; F1-Chi-square with FDR and 10 co-occurrence; P-PRR with 10 co-occurrence; Y-Yule with 10 co-occurrence); top left, top right and bottom are the comparison of precision, recall, and F-measure, respectively.

3.4.3.1.3 Effects of MEDI intervention

In this test, I selected the five statistical models with the best F-measure and evaluated the effects of different MEDI interventions. Since I excluded indications from false pairs in the reference standard (Table 3-3), it is expected that this will decrease the number of false positives (consequently increasing the true negatives). However, the number of true positives will not be changed, so recall should not be changed either. The precision may be improved because of the decreasing of false positives, as may the F-measure. The results
of these experiments are shown in Figure 3-5 and 3-6 below, which present the filtering the indications can decrease the predicted false positives and correspondingly increase the precision and F-measure.

**Table 3-3: Number of false pairs that have been excluded as indications**

<table>
<thead>
<tr>
<th>Reference Set</th>
<th>true pairs</th>
<th>false pairs</th>
<th>Total pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without filtering out indications</td>
<td>71,753</td>
<td>1,696,706</td>
<td>1,768,459</td>
</tr>
<tr>
<td>Filtering indication by MEDI-HPS</td>
<td>71,753</td>
<td>1,694,739</td>
<td>1,766,492</td>
</tr>
<tr>
<td>Filtering indication by MEDI-Complete</td>
<td>71,753</td>
<td>1,690,353</td>
<td>1,762,106</td>
</tr>
</tbody>
</table>

**Figure 3-5:** Comparison of the predicted drug/ADR association count with different MEDI filtering indications for best five performing statistical models (F2-Chi-square with FDR and 25 co-occurrence; C-Chi-square with 25 co-occurrence; F1-Chi-square
with FDR and 10 co-occurrence; P-PRR with 10 co-occurrence; Y-Yule with 10 co-occurrence); top left, top right, bottom left, and bottom right are the comparison of number of true positives, false positives, true negatives, and false negatives, respectively.

**Figure 3-6:** Comparison of ability of different performance metrics to recover SIDER2 side effects for best 5 performing statistical algorithms (F2-Chi-square with FDR and 25 co-occurrence; C-Chi-square with 25 co-occurrence; F1-Chi-square with FDR and 10 co-occurrence; P-PRR with 10 co-occurrence; Y-Yule with 10 co-occurrence) after filtering indications by different MEDI lists; top left, top right and bottom are the comparison of precision, recall, and F-measure, respectively.

### 3.5 Discussion

This experiment demonstrated that unstructured outpatient data can be used as a data source for detecting drug/ADR associations. The statistical testing procedure is well established for signal detection and can be potentially improved by considering the error controlling...
within multiple comparison setting. Filtering known therapeutic relationships from possible drug/ADR candidates can improve the overall performance; this should become a required component in pharmacovigilance using EHR data.

This work has several limitations. MedLEE was used to extract medications and problems from the CDW EHR system. After reviewing the annotated problems, there were some unlikely problem terms such as “arms” or “legs”. These are less likely problems related to medications and maybe caused by NLP processing errors. Since the procedure is not tied to a specific NLP tool, there is also the possibility of selecting another NLP tool with better performance. With the improvement of NLP technology, this method should get better performance accordingly.

There is no agreed-upon gold standard in the field of pharmacovigilance (Tatonetti et al., 2012), and it has been pointed out that SIDER2 data set may contain false drug/ADR associations (Duke J et al., 2011). This is likely to affect the performance evaluation by including wrong validation data. In addition, SIDER doesn’t include all the drugs that were used in the EHR system. Possible causes include the usage of different drug names, new drugs being used in clinical practice and terminology mapping or NLP annotation errors. This further hints at the fact that a comprehensive and accurate reference set is needed for accurately evaluating pharmacovigilance systems or methods.

By exploiting UMLS hierarchical relationships, sub-concepts for a SIDER problem can be grouped together and the pathways between those sub-concepts can be used to relate to the comprising SIDER2 problem concept. However, initial experiments didn’t show an improvement of the overall performance in the experiment. This does not however necessarily imply that grouping related concepts isn’t suitable for improving the signal
detection performance. To assess this better, future studies are needed to scrutinize the effects of subsumption and grouping of concepts that are related through a hierarchy. Ontological inference experiments should be used to group related concepts in a more controlled and accurate manner to determine the effects on signal detection performance.

3.6 Conclusion

Drug/ADR signals can be detected from the outpatient unstructured EHR data with existing disproportionality measures and Chi-square statistics with FDR. Filtering known therapeutic relationships from possible drug/ADR combinations can exclude confounding factors and improve the performance. With the precision of about 10-15% of evaluated SIDER2 by statistical algorithms, additional signal evaluation is needed for the detected drug/ADR signals.
Chapter 4: Using Knowledge Extracted from the Literature to Identify Plausible Adverse Drug Reactions

Over the last decade, drug safety data obtained from spontaneous reporting systems (SRSs) have been analyzed using quantitative data mining procedures to retrieve strongly associated drug/ADR pairs (Puijenbroek et al., 2003; Rawlins, 1988a, 1988b). These highlighted associations are subsequently reviewed and scrutinized by domain experts. Review by domain experts is required to evaluate a signal using their knowledge and judgment to find a signal with clinical significance. However, on account of the human-intensive nature of this task, automated assistance is desirable. In this study, I attempt to partially automate this aspect of the signal evaluation process. I do so using methods that leverage knowledge extracted from the biomedical literature as a means to assess the plausibility of an observed drug/ADR association.

4.1 Materials

In this study, MetaMapped Medline Baseline (MMB) and Semantic MEDLINE Database (SemMedDB) were used to represent knowledge from the biomedical literature. Side Effect Resource 2 (SIDER2) was used as data set for drug/ADR associations. The Semantic Vectors package was used to build concept-based (RRI, Reflective Random Indexing) and

---

predication-based (PSI, Predication-based Semantic Indexing) semantic space models (Cohen et al., 2010; Cohen, Widdows, Schvaneveldt, Davies, et al., 2012).

4.1.1 MMB and SemMedDB

2012 MMB was used as a repository for concept-based modeling. The MMB contains 20,494,848 articles included in Medline up to November, 2011 and contains 399,701 distinct concepts. The SemMedDB V2.2 (semmedVER22) was used for predication modeling, which was processed with SemRep version 1.5. This was the current version when the experiments started. SemMedDB contains 22,252,812 citations included in Medline up to March 31, 2013 and contains 63,795,467 predications. There are 58 distinct predicates and 257,350 distinct concepts in SemMedDB. There are also negated predications in the SemMedDB repository (e.g. anticoagulant_therapy NEG_TREATS (does not TREAT) phlebitis). However, the number of negative predications is relatively small (1.2% of total predications) and consequently they were not included in the PSI model.

4.1.2 SIDER2

SIDER2 is a publicly available database containing information on marketed medicines and their known adverse reactions (Kuhn et al., 2010). SIDER2 was used to construct a dataset for the experiment and as a reference standard to confirm whether a predicted side effect is a true adverse reaction. SIDER2 terms were normalized by mapping drug and side effects terms to UMLS CUI with UMLS Terminology Services (UTS) API 2.0 (U.S. National Library of Medicine & National Institutes of Health, 2012). These UMLS CUIs were then subsequently searched in SemMedDB and MMB to retrieve the mapped UMLS concepts which are represented in SemMedDB and MMB.
SIDER2 contains 996 drugs, 4,192 side effects, and 99,423 drug/ADR pairs. Only those side effects and drugs that were represented in both the RRI and the PSI spaces were retained, so the reference set contains 959 drugs, 3,436 side effects, and 90,787 drug/ADR pairs. Each vector model’s search space was composed of vectors representing the SIDER2 side effects. For the PSI model, SIDER2 drug/ADR pairs were also used as training data to infer predicate reasoning pathways.

4.1.3 MEDication-Indication (MEDI)

As I would anticipate connections in the literature between medications and diseases they treat, I evaluated the utility of another knowledge resource, MEDI (Wei et al., 2013), as a means to eliminate drug indications from consideration as potential side effects. MEDI is a medication indication resource that was extracted from a set of commonly used medication resources, including RxNorm, MedlinePlus, SIDER2, and Wikipedia (Wei et al., 2013). MEDI drugs are represented by RxNorm codes, and indications are represented by ICD-9 codes. MEDI contains 3,112 medications and 63,343 medication-indication pairs. Additionally, the MEDI high-precision subset (MEDI-HPS) was created by only including indications that are retrieved from RxNorm or at least two of the three other resources. MEDI-HPS contains 2,136 medications and 13,304 medication-indication pairs. The estimated precision of MEDI-HPS is about 92% (Wei et al., 2013).

In the experiments, MEDI was used to eliminate drugs’ indications from the side effects search space. To do so, all terms representing drugs (RxCUI) and indications (ICD-9 codes) in MEDI were normalized to UMLS concepts, and then each drug’s indications were filtered from this drug’s search space. In many cases, there exist hierarchical relationships between concepts. For example, C0264702 acute myocardial infarction of apical-lateral
wall is a child node of \textit{C0155626 acute myocardial infarction}. So in the experiments, I extended the MEDI list by aggregating the related concepts by different hierarchical relations. I tested these various extensions of the MEDI list as different MEDI interventions.

4.2 Methods
4.2.1 RRI

RRI (Cohen et al., 2010) is a variant of RI adapted to enable the recognition of meaningful indirect associations. The variant of RRI I used for the experiments allows for the estimation of semantic relatedness between UMLS concepts, and proceeds as follows. First, all terms in the text corpus are assigned unique vector representations, known as elemental vectors. I will refer to the elemental vector for concept C as $E(C)$ for the remainder of this manuscript. In accordance with the RI paradigm (Kanerva, 2009), elemental vectors are generated stochastically. In this way, RI creates unique fingerprints for all terms in the text corpus. The vector components can be binary, ternary, real, or complex values (Kanerva, 2009; Widdows & Cohen, 2012). In the experiments, I use 32,000 dimensional binary vectors constructed in accordance with the Binary Spatter Code (BSC) (Kanerva, 1994), one of a family of representational approaches known as Vector Symbolic Architectures (VSAs) (Gayler, 2004; Kanerva, 1994, 1996; Plate, 1995). This dimensionality was selected based on the results of simulation experiments in previous research (Wahle, Widdows, Herskovic, Bernstam, & Cohen, 2012), which suggest that at this dimensionality around 2,000 unique elemental vectors can be superposed with low probability of the superposed product being closer to some other elemental vector in the
space than its component vectors. However, I did not attempt to optimize this parameter, and would anticipate some improvement in accuracy in exchange for the additional computational work required to perform these experiments at higher dimensionalities. In the BSC, elemental vectors are constructed by distributing an equal number of 1’s and 0’s at random across the dimensions of the vector concerned. Consequently, elemental vectors have a high probability of being orthogonal or close-to-orthogonal to each other, with orthogonality defined as a Hamming Distance (HD) of half the dimensionality of the vectors concerned (Kanerva et al., 2000; Kanerva, 1994, 1996).

The next step is to generate vector representations of documents, by superposing the elemental vectors of the terms contained in these documents. With binary vectors, superposition is accomplished by keeping track of the number of 1’s and 0’s that have been added in each dimension, and assigning the value in this dimension using the majority rule, with ties split at random. I will refer to this operation by using the “+” symbol, with “+=” indicating a superposition that includes the vector on the left of the operator also (so \( \text{DOC}(D) += \text{E}(C) \) is equivalent to \( \text{DOC}(D) = \text{DOC}(D) + \text{E}(C) \), a common operation during training).

In the experiments, this superposition is weighted using the Log-Entropy weighting procedure. The local term weight for term \( i \) in document \( j \) \((l_{ij})\) is derived from the frequency of a term in a document. The global weight for term \( i \) \((g_i)\) describes the frequency of the term within the entire text corpus. They are computed with Equation 1.

This weighting scheme reduces the influence of high frequency terms that may be uninformative, and tempers the influence of terms that recur frequently within a single document (Dumais, 1991). Once document vectors have been generated (Equation 2), it is
possible to generate vector representations of concepts (in this case, or terms in the general case), known as semantic vectors. I will refer to the semantic vector for concept C as $S(C)$.

Semantic vectors are constructed by superposing vector representations of the documents a concept occurs in.

\[ \text{LogEntropy}(\text{term } i, \text{document } j) = l_{ij} \times g_i \]
\[ l_{ij} = \log(\text{term } i, \text{document } j) = \log(1 + tf_{ij}) \]
\[ g_i = \text{Entropy}(\text{term } i) = 1 - \sum_j p_{ij} \times \frac{\log p_{ij}}{\log n}, \text{while } p_{ij} = \frac{tf_{ij}}{gf_i} \]

$tf_{ij}$: number of occurrences of term $i$ in document $j$

$gf_i$: global frequency for term $i$ (total occurrences of term $i$ in text corpus)

$n$: total number of documents in text corpus

**Equation 1**: Log-Entropy weighting equation

\[ S(\text{doc}) = E(\text{term}) \cdot \text{local weight} \cdot \text{global weight} = E(\text{term}) \cdot l_{ij} \cdot g_i \]

**Equation 2**: Generating document vectors with weighting procedure in RRI

Superposition of binary vectors requires maintaining a “voting record” that keeps track of the number of 1’s and 0’s added in each dimension. When local and global weighting metrics are utilized, the “votes” may not be integer values. So, for example, if the vector 1010 were added with a weight of 0.5, a straightforward implementation of the voting record would add 0.5 to the dimensions of the voting record corresponding to the 1’s, and subtract 0.5 from the dimensions corresponding to the 0’s. Normalization involves tallying these votes. After training is complete, those dimensions of the voting record with positive values would be assigned 1, those with negative values would be assigned 0, and those with a zero value would be assigned either 1 or 0 at random. In practice, however, it is
computationally inconvenient to maintain and update 32,000 real values to serve as a voting record for each semantic vector. Consequently, the Semantic Vectors package employs a binary matrix approximation of the voting record, which sacrifices some floating-point precision in exchange for computational efficiency. These implementation details are provided in (Widdows & Cohen, 2012).

These operations are expressed concisely in the pseudo code in Figure 4-1, adapted from (Cohen et al., 2010). A schematic representation for RRI is shown in Figure 4-2. I used Semantic Vectors Version 3.7 to build RRI vectors. Once semantic vectors were constructed, the relatedness between drugs and ADRs was estimated as \[1 - \frac{2}{n} HammingDistance(x, y)\]. Therefore, a ranked list of ADRs for each drug was provided.

\[
q := \# \text{total terms} \\
n := \# \text{dimensions}, n := 32,000 \\
p := \# \text{total documents} \\
m := \# \text{total UMLS concepts} \\
k := \# \text{terms in a specific document OR \# documents related to a specific concept} \\
\triangleright \text{initialize elemental vectors } E(T) \\
\text{for } i < q \text{ do} \\
\quad \text{generate a zero vector of dimension } n \\
\quad \text{in } E(T_i) \text{ arbitrarily set half of the zero dimensions to 1} \\
\text{end for} \\
\triangleright \text{train document model } (E(T), p) \\
\text{for } j < p \text{ do} \\
\quad \text{for term } i \text{ in document } j \text{ do} \\
\quad\quad \text{DOC}(D_j) += E(T_i) \times \text{LogEntropy} \text{ (term } i, \text{ doc } j) \\
\quad \text{end for} \\
\quad \text{normalize } (\text{DOC} (D_j)) \\
\text{end for}
\]
\[\text{for } j < m \text{ do}\]

\[\text{for each of } k \text{ documents concept } j \text{ occurs in do}\]

\[S(C_j) + = \text{DOC}(D_k)\]

end while

normalize \((S(C_j))\)

end for

**Figure 4-1:** The pseudo code for RRI model training process

**Figure 4-2:** Schematic representation of RRI training and inference process
4.2.2 PSI

4.2.2.1 Operations in PSI

The PSI model provides the means to implement discovery patterns for LBD using distributional semantics (Cohen, Widdows, Schvaneveldt, Davies, et al., 2012; Cohen, Widdows, Schvaneveldt, & Rindflesch, 2012). This is accomplished by representing concepts and relationships extracted by SemRep as high-dimensional vectors using an adaption of RI. In previous work, PSI has been applied to discover therapeutic relationships (Cohen, Widdows, Schvaneveldt, Davies, et al., 2012) using a two-stage process of discovery by analogy: first a geometric operator is used to infer discovery patterns from known treatments, then the identified discovery patterns are used to infer previously unseen therapeutic relationships.

In addition to the superposition operation described previously, the PSI model utilizes a binding operation. Binding (⊗) is a compositional operation that is provided by VSAs, such as the BSC (Gayler, 2004; Kanerva, 1996). Binding two elemental vectors generates a third vector, which is dissimilar from these two component vectors. The binding operation is reversible (release ◯). With binary vectors, pairwise exclusive OR (XOR) is used to accomplish both binding (⊗) and release (◯).

4.2.2.2 PSI training process

The training process for generating semantic vectors proceeds as follows:

(1) Generate elemental vectors for all concepts and relations occurring in semantic predications;

(2) generate a semantic vector for each concept, initially empty;
(3) for each predication (concept-predicate-concept), bind the elemental vector of one concept and the elemental vector of the predicate, and add this bound product to the semantic vector for the other concept.

During step (3), a statistical weighting scheme is applied. For the predication $C_1 P C_2$, the semantic vector $S(C_2)$ is generated as shown in Equation 3.

$$S(C_2) = E(C_1) \cdot Pf \cdot (idf_P + idf_{C_1})$$

Equation 3: Generating second degree of Semantic Vectors

The global weight $Pf$ is derived from the number of times that the predication occurs in the SemMedDB. The local weight $idf$ (inverse document frequency of the concept $C$ or the predicate $P$) reflects the occurrence of the concept across all documents. They are computed as shown in Equation 4.

$$Pf = \log(1 + \text{occurrences of predication } C_1 P C_2)$$

$$idf_{C/P} = \log\left(\frac{\text{number of total predications}}{\text{number of predications containing } C/P}\right)$$

Equation 4: Weighting procedure in PSI

The pseudo code for PSI is displayed in Figure 4-3.

$q := \# \text{ unique UMLS concepts and predicates from SemMedDB}$

$n := \# \text{ dimensions, } n := 32,000$

$p := \# \text{ total predications}$

$k := \# \text{ predications related to a specific concept}$

$\triangleright$ initialize elemental vectors $E(C)$ and $E(P)$

for $i < q$ do
generate a zero vector of dimension $n$
in $E(C_1)$ or $E(P_i)$ arbitrarily set half of the zero dimensions to 1
\end for
\begin{itemize}
\item[] \textbf{\triangleright} train semantic model $(E, p)$
\item[] for $j < p$ do
\item[] predication $j$: $C_1 P C_2$
\item[] $S(C_1) += E(C_2) \otimes E(P - INV) \cdot Pf \cdot (idf_P + idf_{C_2})$
\item[] $S(C_2) += E(C_1) \otimes E(P) \cdot Pf \cdot (idf_P + idf_{C_1})$
\end for
\end{itemize}
\begin{itemize}
\item[] normalize $S(C)$
\item[] \textbf{Figure 4-3}: The Pseudo code for PSI model training process
\end{itemize}

All concepts and relations were assigned a binary elemental vector of 32,000 bits in length. The semantic vector of each concept was generated by superposing bound products related to this concept, where the bound products were produced by binding the elemental vectors for the other concept and predicate elemental vectors in each predication this concept occurs in. The search space of SIDER2 side effects contains 3,436 ADRs.

\subsection{4.2.2.3 Inferring discovery patterns}
After training the semantic vectors, the PSI model can be used to infer discovery patterns by "releasing" the semantic vector of a drug using the semantic vector of its ADR. The bound product of the drug’s semantic vector and discovery patterns’ vectors can be subsequently used as a query vector to search the vector space of side effects. In this procedure, discovery patterns were inferred from all known drug/ADR associations. For each drug, the five discovery patterns that were most frequently inferred from all other drugs and their ADRs were retained.
The pathways connecting drugs to side effects may not be restricted to one middle term (and two predicates). In previous experiments predicting therapeutic relationships, performance was improved by including pathways of three predicates and two middle terms (Cohen, Widdows, Schvaneveldt, & Rindflesch, 2012). This is accomplished by generating a second-degree semantic vector for a concept, \( S_2(\text{concept}) \), by adding together the (first-degree) semantic vectors of all concepts connected to it by a predicate of interest. In the experiments, the two most popular predicates from inferred double-predicate reasoning pathways -- INTERACTS_WITH and COMPARED_WITH -- were used to build second-degree semantic vectors \( S_2 \). This vector is then used as an alternative starting point for the inference procedure. From this point, the five most frequently inferred double-predicate reasoning pathways using the second order semantic vector of all other drugs and the (first order) semantic vectors of their ADRs were retained. As these inferred pathways connect to drugs through either INTERACTS_WITH or COMPARED_WITH, they are referred to as triple-predicate pathways.

### 4.2.2.4 Applying discovery patterns to find possible ADRs (Step 5 in Figure 4-4)

To combine query vectors for frequently inferred reasoning pathways into one search expression, I use a disjunction operation that originates in the quantum logic of Birkhoff and von Neumann, and was first applied to information retrieval by Widdows and Peters (Birkhoff & Von Neumann, 1936; Widdows & Peters, 2003). I define the disjunction of these five query vectors as a query subspace derived from them using a binary vector approximation (Cohen, Widdows, De Vine, Schvaneveldt, & Rindflesch, 2012) of the Gram-Schmidt orthonormalization procedure (Golub & Van Loan, 2012). The length of
the projection of some other vector in this subspace provides an estimate of vector-subspace similarity.

For the double-predicate discovery patterns model, a drug’s query subspace was constructed from this drug’s first-degree semantic vector bound to the vector representations of the five double predicate reasoning pathways most frequently inferred from other drugs. For the double- and triple-predicate discovery patterns model, a drug’s query subspace also included this drug’s second-degree semantic vector bound to vector representations of the five reasoning pathways most frequently inferred from the second-degree semantic vectors of other drugs.

The length of the projection of the semantic vector for a candidate ADR into a drug’s query subspace was used to estimate the relatedness between these entities, providing a ranked list of potential ADRs for each drug.

Figure 4-4 provides an overview of the PSI-based analogical reasoning process in its entirety.
Step 1: Randomization of Concepts and Predicates
Elemental vectors: $E(\text{Rosiglitazone})$, $E(\text{IW})$, $E(\text{TG})$, $E(\text{AW})$, $E(\text{MI})$, $E(\text{SomeConcept})$...

Step 2: Training Semantic Vectors for Concepts
By superposition of related bound products with weighting procedure

\[
S(\text{rosiglitazone}) + = E(\text{IW}) \otimes E(\text{TG}) \times $ \text{Pf-idf} \\
+ = ... \\
S(\text{MI}) + = E(\text{TG}) \otimes E(\text{AW}) \times $ \text{Pf-idf} \\
+ = ...
\]

Step 3: Infer Double Predicate Discovery Patterns
By releasing $S(\text{ADR})$ from $S(\text{drug})$ for all SIDER2 drug/ADR pairs

\[
S(\text{rosiglitazone}) \otimes S(\text{MI}) \approx E(\text{IW}) \otimes E(\text{TG}) \otimes (E(\text{TG}) \otimes E(\text{AW})) \\
\approx E(\text{IW}) \otimes E(\text{TG}) \otimes E(\text{TG}) \otimes E(\text{AW})) \\
\approx E(\text{IW}) \otimes E(\text{AW})...
\]

Step 4: Build Second Degree Semantic Vectors of Drugs and Infer Triple Predicate Discovery Patterns
$C_2$ are SomeConcept that IW or CoeW with rosiglitazone

\[
S_2(\text{rosiglitazone}) = \sum C_2 \\
S_2(\text{rosiglitazone}) \otimes S(\text{MI}) \approx E(\text{CoeW}) \otimes E(\text{AW})
\]

Step 5: Build Query Subspace as Search Vector to Infer ADR

\[
S(\text{rosiglitazone}) \otimes (E(\text{IW}) \otimes E(\text{CAUSES-INV})) \\
\lor S(\text{rosiglitazone}) \otimes (E(\text{ISA}) \otimes E(\text{CAUSES-INV})) \\
\lor S(\text{rosiglitazone}) \otimes (E(\text{AW}) \otimes E(\text{IW})) \\
\lor S(\text{rosiglitazone}) \otimes (E(\text{AW}) \otimes E(\text{CoeW})) \\
\lor S(\text{rosiglitazone}) \otimes (E(\text{ComW}) \otimes E(\text{CAUSES-INV})) \\
\lor S_2(\text{rosiglitazone}) \otimes (E(\text{IW}) \otimes (E(\text{CAUSES-INV})) \\
\lor S_2(\text{rosiglitazone}) \otimes (E(\text{AW}) \otimes E(\text{AFFECTS-INV})) \\
\lor S_2(\text{rosiglitazone}) \otimes (E(\text{AW}) \otimes E(\text{CoeW})) \\
\lor S_2(\text{rosiglitazone}) \otimes (E(\text{ComW}) \otimes E(\text{CAUSES-INV})) \\
\approx S(\text{ADR})
\]

Figure 4-4: Schematic representation of PSI training and inference process.
Triglycerides: TG; myocardial infarction: MI; INTERACTS_WITH: IW; COEXISTS_WITH: CoeW; ASSOCIATED_WITH: AW; COMPARED_WITH: ComW
4.3 Experimental design

An overview of the experimental design is shown in Figure 4-5. The first experiment was conducted without knowledge of drug indications. The concept-based RRI model and discovery pattern-based PSI model were compared with respect to their ability to identify known drug/ADR associations. In the second experiment, the model with the best performance from the first experiment was used to evaluate the effect of eliminating known indications from the list of predictions.

**Figure 4-5:** Experimental design in the detection of SIDER2 known ADRs using LBD distributional semantic models
4.3.1 Experiment 1 design

Distributional semantic vectors were used to model MMB and SemMedDB. RRI vectors and PSI vectors formed the basis for the models of LBD concept-based co-occurrence and LBD discovery patterns, respectively. As MetaMap may retrieve many more concepts from a particular document than SemRep retrieves predications, I varied the RRI model to assess the extent to which observed effects were due to the advantage of a more extensive (albeit less structured) knowledge base. In one case, a RRI space was derived from only those sentences from which predications were extracted. Consequently, there are three distributional semantic models -- RRI built from documents (RRI-from-document group), RRI built from predication source sentences (RRI-from-predication group), and PSI built from predications. The PSI model was evaluated with two settings. In the first case, only two-predicate discovery patterns were considered (PSI-double group), while the second case considered both two- and three-predicate patterns (PSI-double+triple group). The elemental vectors for terms, which are not meaningfully related to one another, were used to implement a random baseline (Baseline group).

With the RRI models, for each drug, related problems were sought by comparing each vector in the side effect search space to this drug’s vector representation.

With the PSI model, SIDER2 known drug-side effect pairs were used to infer predicate paths. For the PSI-double group, each drug’s query subspace was built as the disjunction of the bound products between the drug and its five double-predicate reasoning pathways. For the PSI-double+triple group, each drug’s query subspace additionally included the second degree semantic vector of this drug bound to five triple-predicate paths. The five triple-predicate paths were retrieved by the extension of second degree semantic vectors of
drugs. Comparing a drug’s query subspace with each vector in the search space allowed us to infer the drug’s possible side effects.

4.3.2 Experiment 2 design

From preliminary results, I found that there were some indications in the inferred ADRs. So I hypothesized that excluding known indications for drugs from the search space would improve performance. I tested this hypothesis in the second experiment utilizing knowledge of drug indications from MEDI. In this experiment, I tested variants of the MEDI indication list using the best performing model from the first experiment (PSI double + triple group). I extended the MEDI-complete and MEDI-HPS lists to include all offspring, or immediate offspring nodes based on the UMLS semantic network utilizing the MRREL.RRF file. This file includes relationships between UMLS concepts found in the UMLS Metathesaurus (U.S. National Library of Medicine, 2009). By utilizing these ancestor-offspring hierarchical relationships, I define an offspring node as a node that has a MEDI indication as an ancestor (regardless of the number of intervening nodes); and an immediate offspring node as a node that has this MEDI indication as its parent.

In this procedure, I first normalized all MEDI terms. For MEDI drugs, I mapped each drug’s RxCUI to a UMLS CUI with the RxNorm API (U.S. National Library of Medicine & National Institutes of Health, 2014) and then searched the UMLS CUI in SemMedDB and MMB to retrieve the mapped UMLS concept. For MEDI indications, I mapped each indication’s ICD-9 term to a UMLS CUI using the UTS API 2.0 (U.S. National Library of Medicine & National Institutes of Health, 2012) and then subsequently searched for this UMLS CUI in SemMedDB and MMB to retrieve the mapped UMLS concept. After
normalizing MEDI terms, the hierarchical relation of synonym (SY), child (CHD), and sibling (SIB) in MRREL.RRF were used to find drugs’ MEDI indications extended offspring or immediate offspring. Consequently, there were six MEDI lists (Table 4-1). These MEDI lists were used to exclude indications from the side effect search space and were tested in the second experiment.

<table>
<thead>
<tr>
<th>MEDI intervention</th>
<th>Extension Procedure</th>
<th>Experiment Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>No MEDI</td>
<td>None</td>
<td>No-MEDI group</td>
</tr>
<tr>
<td>MEDI-HPS indications for SIDER2 drugs</td>
<td>None</td>
<td>MEDI-HPS group</td>
</tr>
<tr>
<td></td>
<td>All synonym (SY), child (CHD), and sibling (SIB), as well as their offspring.</td>
<td>MEDI-HPS-offspring group</td>
</tr>
<tr>
<td></td>
<td>Immediate synonym (SY), child (CHD), and sibling (SIB) relationships.</td>
<td>MEDI-HPS-immediate offspring group</td>
</tr>
<tr>
<td>MEDI-complete indications for SIDER2 drugs</td>
<td>None</td>
<td>MEDI-complete group</td>
</tr>
<tr>
<td></td>
<td>All synonym (SY), child (CHD), and sibling (SIB), as well as their offspring.</td>
<td>MEDI-complete-offspring group</td>
</tr>
<tr>
<td></td>
<td>Immediate synonym (SY), child (CHD), and sibling (SIB) relationships.</td>
<td>MEDI-complete-immediate offspring group</td>
</tr>
</tbody>
</table>

4.3.3 Performance measurements

To evaluate performance, I used a number of widely used metrics. Precision measures the proportion of accurate ADRs in relation to the total number of ADRs retrieved (Salton, Fox, & Wu, 1983). To evaluate the precision at different points in a ranked list, I used average precision (AP, the average of the precision values measured at the point at which
each correct result is retrieved for one example (Manning, Raghavan, & Schütze, 2008)). Mean average precision (MAP) is the average of the AP across all drugs. Precision at $k$ (Manning et al., 2008) measures the precision at fixed levels of retrieved results and emphasizes the importance of finding relevant results early. I evaluated precision at $k=50$ ($P_{k=50}$). Recall represents the proportion of ADRs retrieved out of the total number of ADR associations in the reference standard (Salton et al., 1983).

A “rediscovery” (true discovery) is defined as an adverse effect inferred by a vector model and subsequently confirmed by SIDER2 as a true prediction. Consequently, the median rediscovery rank for a particular drug approximates the point in the ranked list produced by a particular model at which half of the known adverse reactions for this drug were recovered.

The AP and median rank of the rediscoveries across drugs were compared by the paired $t$ test and the Wilcoxon matched-pairs signed-rank test, respectively.

To measure the performance with respect to the true positive rate (TPR) and false positive rate (FPR), receiver operating characteristic (ROC) curve was plotted for all drug/ADR pairs for all models. Subsequently, a global area under the ROC curve (AUC, “global” indicates that the scores of all drug/ADR pairs were combined into a single curve) was calculated using AUCCalculator (Davis & Goadrich, 2006). For the model with the best global AUC, a drug-based AUC was also calculated and compared between drugs.

4.4 Results

4.4.1 Experiment 1
4.4.1.1 Inferring discovery patterns

The most strongly associated double-predicate path was calculated for each known drug/ADR pair. In total, 90,787 predicate paths were inferred. Among them, there were 1,485 unique predicate paths. The five most frequently inferred double-predicate paths were selected. Second degree semantic vectors for drugs were constructed by adding together the semantic vector representations of any concept occurring in a semantic predication with the drug in question, where the predicate type was either INTERACTS_WITH or COMPARED_WITH. The most frequently occurring double predicate paths and inferred triple predicate paths with corresponding examples are shown in Table 4-2. They are consistent across all drugs. Many of these paths are readily interpretable, and could support a plausible biological mechanism for a predicted effect. For example, INTERACTS_WITH:CAUSES-INV suggests a drug may interfere with some biological factor which may cause a side effect. COMPARED_WITH:CAUSES-INV can be used to identify similar side effects by comparing their drug class information as COMPARED_WITH often indicates a comparative evaluation across different drugs in the same therapeutic category. Triple predicate paths extend the connecting path for drugs and related ADRs.

Table 4-2: The most frequent predicate paths in inferred discovery patterns

<table>
<thead>
<tr>
<th>double/triple-predicate</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERACTS_WITH CAUSES-INV</td>
<td>dipyridamole INTERACTS_WITH nitric oxide</td>
</tr>
<tr>
<td></td>
<td>bradycardia CAUSES-INV nitric oxide</td>
</tr>
</tbody>
</table>

74
<table>
<thead>
<tr>
<th>double/triple-predicate</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASSOCIATED_WITH COEXISTS_WITH</td>
<td>rosiglitazone COEXISTS_WITH apolipoprotein a-ii apolipoprotein a-ii ASSOCIATED_WITH myocardial infarction</td>
</tr>
<tr>
<td>COMPARED_WITH CAUSES-INV</td>
<td>bisoprolol COMPARED_WITH metoprolol hypotension CAUSES-INV metoprolol</td>
</tr>
<tr>
<td>ASSOCIATED_WITH INTERACTS_WITH</td>
<td>rosiglitazone INTERACTS_WITH triglycerides triglycerides ASSOCIATED_WITH myocardial infarction</td>
</tr>
<tr>
<td>ISA CAUSES-INV</td>
<td>naproxen ISA calcineurin inhibitor toxic nephropathy CAUSES-INV calcineurin inhibitor</td>
</tr>
<tr>
<td>INTERACTS_WITH INTERACTS_WITH ASSOCIATED_WITH</td>
<td>rosiglitazone INTERACTS_WITH lyrm1 lyrm1 INTERACTS_WITH fatty acids, nonesterified fatty acids, nonesterified ASSOCIATED_WITH myocardial infarction</td>
</tr>
<tr>
<td>INTERACTS_WITH ASSOCIATED_WITH COEXISTS_WITH</td>
<td>rosiglitazone INTERACTS_WITH glycerol-3-phosphate dehydrogenase glycerol-3-phosphate dehydrogenase COEXISTS_WITH succinate dehydrogenase Succinate dehydrogenase ASSOCIATED_WITH myocardial infarction</td>
</tr>
<tr>
<td>COMPARED_WITH INTERACTS_WITH ASSOCIATED_WITH</td>
<td>rosiglitazone COMPARED_WITH glycerophosphates glycerophosphates INTERACTS_WITH low-density lipoproteins low-density lipoproteins ASSOCIATED_WITH myocardial infarction</td>
</tr>
<tr>
<td>COMPARED_WITH COEXISTS_WITH ASSOCIATED_WITH</td>
<td>rosiglitazone COMPARED_WITH gw 501516 gw 501516 COEXISTS_WITH high-density lipoprotein cholesterol high-density lipoprotein cholesterol ASSOCIATED_WITH myocardial infarction</td>
</tr>
</tbody>
</table>
4.4.1.2 Performance

Results for different vector models are shown in Table 4-3. PSI-based models performed better than RRI-based models and both models perform better than the random baseline. The PSI-double + triple group outperformed all other groups. All differences in median rank and AP were statistically significant (as estimated by Wilcoxon’s signed rank test and paired t test respectively). $P_{k=50}$ for each drug was compared across groups using Pearson’s correlation. For variants of the same model (RRI or PSI), $P_{k=50}$ was highly correlated (0.75-0.84). Correlation in $P_{k=50}$ between the PSI and RRI models was between 0.52 and 0.57, suggesting the potential to improve performance by combining results.

Table 4-3: Results of precision and rank-based measures for different groups

<table>
<thead>
<tr>
<th>Group</th>
<th>MAP</th>
<th>$P_{k=50}$ for all drugs (global precision)</th>
<th>Median Rank (n=3,436)</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(global precision)</td>
<td>median</td>
<td>mean</td>
</tr>
<tr>
<td>Baseline group</td>
<td>0.0300</td>
<td>0.0284</td>
<td>1708</td>
<td>1711.44</td>
</tr>
<tr>
<td>RRI-from-predication group</td>
<td>0.0365</td>
<td>0.0469</td>
<td>1629</td>
<td>1651.08</td>
</tr>
<tr>
<td>RRI-from-document group</td>
<td>0.0520</td>
<td>0.0784</td>
<td>1333</td>
<td>1454.30</td>
</tr>
<tr>
<td>PSI-double group</td>
<td>0.0591</td>
<td>0.0942</td>
<td>1233</td>
<td>1379.65</td>
</tr>
<tr>
<td>PSI-double+triple group</td>
<td>0.0848</td>
<td>0.1410</td>
<td>808</td>
<td>1108.47</td>
</tr>
</tbody>
</table>

4.4.1.3 AUC

Figure 4-6 and Table 4-3 present the global ROC curves for all models. ROC curve shows the tradeoff between sensitivity and specificity. The global AUC provides a cumulative estimate of accuracy, and is shown for each model in Table 4-3. PSI-double+triple group
has the best global AUC of 0.6841. I measured its AUC at the drug level (Figure 4-7). The mean and median AUC are 0.7102 ± 0.0752 and 0.7058 respectively. Figure 4-7 shows a plot of the AUC for each drug against the log of the number of predictions in SemMedDB with this drug as subject. This suggests a trend in which performance is generally better for those drugs for which more knowledge is available in the database. Those drugs with an AUC of 0.8 or above tend to occur in 10,000 or more predications as subject.

Note that the global AUC as a metric will be inflated by methods that incorporate category bias into their prediction (Akbani, Kwek, & Japkowicz, 2004; Liu, Wu, et al., 2012), a subject I will return to in the discussion section.

![ROC plot of true positive rate and false positive rate for all groups](image)

**Figure 4-6:** ROC plot of true positive rate and false positive rate for all groups
4.4.1.4 Rediscovery results

The results of this experiment are illustrated in Figure 4-8. This figure plots the number of rediscovered side effects (left Y axis) and the proportion of the valid side effects rediscovered (or global recall, right Y axis) for each model against the mean number of suggested potential ADRs (X axis) at different statistical thresholds. All distributional models outperform the random baseline.
Figure 4-8: Rediscovery plot for experiment groups

With approximately 100 predictions per drug, baseline, RRI-from-predication, RRI-from-document, PSI-double and PSI-double+triple group have a global recall of 0.029, 0.045, 0.069, 0.088, 0.125, respectively.

4.4.2 Experiment 2

The PSI-double+triple model was the best performing model in the first experiment, and was selected to test the effects of using variants of the MEDI list as a way to exclude therapeutic relationships to reduce the number of highly ranked false positive predictions. Table 4-4 presents the performance of the PSI-double+triple model when different MEDI lists were used. The median rank of true positive predictions was lower when MEDI was used to exclude the indication from the search space for each drug. However, as median
rank is based on the rank of true positive results only, it does not consider known side
effects that may have been excluded from consideration by the MEDI list. In contrast, MAP
also measures whether true side effects have been excluded. Consequently, MAP in the
MEDI-complete-immediate offspring was higher than other groups. Overall, AUC was
highest for the MEDI-HPS-immediate offspring group. Of the models, only the MEDI-
HPS-immediate offspring group outperformed the baseline PSI model by all metrics, and
the improvements in performance were small in comparison with the differences in
performance between distributional models in experiment 1. All differences between all
MEDI intervention groups and No-MEDI group in $P_{k=50}$ are statistically significant as
measured by the paired $t$ test. However, the improvement in cumulative accuracy is
negligible.

Table 4-4: *PSI-double+triple model performance across all tests with different MEDI
lists. Best results are in boldface. Performance exceeding the baseline (results obtained
by the best PSI model without MEDI) is marked with an asterisk (*)*

<table>
<thead>
<tr>
<th>Group</th>
<th>MAP</th>
<th>$P_{k=50}$ for all drugs</th>
<th>Median Rank</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>median</td>
<td>mean</td>
</tr>
<tr>
<td>No-MEDI group</td>
<td>0.0848</td>
<td>0.1410</td>
<td>808</td>
<td>1108.468</td>
</tr>
<tr>
<td>MEDI-HPS group</td>
<td>0.0849*</td>
<td>0.1417*</td>
<td>807*</td>
<td>1107.595*</td>
</tr>
<tr>
<td>MEDI-HPS-offspring group</td>
<td>0.0798</td>
<td>0.1283</td>
<td>765*</td>
<td>1050.745*</td>
</tr>
<tr>
<td>MEDI-HPS-immediate offspring group</td>
<td>0.0866*</td>
<td>0.1444*</td>
<td>795*</td>
<td>1094.41*</td>
</tr>
<tr>
<td>MEDI-complete group</td>
<td>0.0839</td>
<td>0.1401</td>
<td>809</td>
<td>1108.062*</td>
</tr>
<tr>
<td>MEDI-complete-offspring group</td>
<td>0.0673</td>
<td>0.0917</td>
<td>681*</td>
<td>953.761*</td>
</tr>
<tr>
<td>MEDI-complete-immediate offspring group</td>
<td>0.0889*</td>
<td>0.1467*</td>
<td>773*</td>
<td>1068.718*</td>
</tr>
</tbody>
</table>
4.4.3 Plausibility evidence found by PSI discovery patterns approach

In this paper, the association between rosiglitazone and myocardial infarction, a highly publicized ADR discovered after the drug was released to the market, is used to illustrate how evidence from the literature can be retrieved for the evaluation of plausibility by a domain expert. The term “myocardial_infarction” was ranked in the top 1% (rank=29) and top 1.5% (rank=50) of potential side effects for rosiglitazone by the PSI-double and PSI-double+triple models respectively.

Rosiglitazone is a thiazolidinedione (TZD) antidiabetic drug, used to treat type 2 diabetes mellitus as an adjunct to lifestyle changes (Cheung, 2010; Hamblin, Chang, Fan, Zhang, & Chen, 2009; Lygate et al., 2003). Since its approval by the FDA in 1999, rosiglitazone was prescribed 3.8 million times annually up to June 2009 in the United States (Shah et al., 2010). A meta-analysis of clinical trials conducted by Nissen (Nissen & Wolski, 2007) in 2007 suggested that the use of rosiglitazone was associated with a significant increase in the risk of myocardial infarction. This led to rosiglitazone’s withdrawal from the European market in 2010 and a rosiglitazone black-box warning in the U.S. (Berthet, Olivier, Montastruc, & Lapeyre-Mestre, 2011; Shah et al., 2010). In 2013, the FDA lifted some prescription restrictions in the U.S. market based on a reevaluation of the Rosiglitazone Evaluated for Cardiac Outcomes and Regulation of Glycaemia in Diabetes (RECORD) trial (ClinicalTrials.gov Identifier NCT00379769) (U.S. Food and Drug Administration, 2013), but the European suspension is still in effect at the time of this writing.

Rosiglitazone is a nuclear peroxisome proliferator-activated receptor (PPAR-gamma) agonist. The mechanism through which rosiglitazone causes cardiovascular events is unclear, but is thought to be related to unfavorable effects on triglycerides, low-density lipoprotein cholesterol, and plasma renin activity.
lipoprotein cholesterol (LDL-C) particle size and density, and greater affinity for PPAR-
gamma than other TZD drugs (Bourg & Phillips, 2012; Goldberg et al., 2005; Khanderia,
Pop-Busui, & Eagle, 2008; Y. K. Loke et al., 2011). To evaluate the extent to which these
hypotheses were consistent with information utilized by the PSI-double+triple model, I
reconstructed the pathways of predicates and concepts that were consistent with the
inferred discovery patterns used to make this prediction.

For myocardial infarction, each discovery pattern that was used for the inference was used
to search the indexed SemMedDB predications and find middle terms that connect
rosiglitazone with myocardial infarction through the discovery pattern. The middle terms
retrieved were ranked based on their inverse document frequency. Since the indexed
SemMedDB predications contain the source literature ID (PMID), I also retrieved related
literature evidence that supports the prediction.

Consequently 108,100 unique predication pathways were retrieved through 8 unique
predicate paths (Table 4-5) and with 1,601 distinct middle terms that connect rosiglitazone
with myocardial infarction. Table 4-6 shows some example predication pathways that were
composed of two or three predications. There were around 17 sentences providing evidence
to support each predication on average. I analyzed middle terms’ semantic groups
(Bodenreider & McCray, 2003) and list the sample with distinct predicate paths connecting
with different semantic groups (Figure 4-9).

Table 4-5: Reasoning pathways used to retrieve evidence from the literature for the pair
rosiglitazone -- myocardial infarction

<table>
<thead>
<tr>
<th>Predicate Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPARED_WITH : COEXISTS_WITH : ASSOCIATED_WITH</td>
</tr>
<tr>
<td>COMPARED_WITH : INTERACTS_WITH : ASSOCIATED_WITH</td>
</tr>
</tbody>
</table>

82
Table 4-6: Some example predications for possible mechanism of rosiglitazone causing myocardial infarction

Middle term “LDL cholesterol lipoprotein”; 123 unique predication pathways; for example:
rosiglitazone INTERACTS_WITH apolipoproteins_b (Brackenridge et al., 2009; Sarafidis et al., 2005) INTERACTS_WITH ldl_cholesterol_lipoproteins (Vessby, Kostner, Lithell, & Thomis, 1982) ASSOCIATED_WITH myocardial_infarction (Goldberg et al., 2005)
rosiglitazone INTERACTS_WITH paraoxonase_1 (van Wijk et al., 2006) INTERACTS_WITH ldl_cholesterol_lipoproteins (Gupta, Singh, Maturu, Sharma, & Gill, 2011) ASSOCIATED_WITH myocardial_infarction (Tetsuro Yoshida et al., 2010)

Middle term “triglyceride”; 1515 unique predication pathways; for example:
rosiglitazone COEXISTS_WITH triglycerides (Goldberg et al., 2005) ASSOCIATED_WITH myocardial_infarction (De Caterina et al., 2011; Friis-Moller et al., 2003; Kuller Lewis et al., 2002; Lekhal, Børvik, Brodin, Nordøy, & Hansen, 2010; Phillips, 1977)
rosiglitazone INTERACTS_WITH triglycerides (Chao et al., 2000; Nadeau, Ehlers, Aguirre, Reusch, & Draznin, 2007) ASSOCIATED_WITH myocardial_infarction rosiglitazone INTERACTS_WITH glycerol-3-phosphate_dehydrogenase (Suzuki, Suzuki, Sembon, Fuchimoto, & Onishi, 2013) COEXISTS_WITH triglycerides (Im & Hoopes, 1983) ASSOCIATED_WITH myocardial_infarction

Middle term “ppar_gamma”; 992 unique predication pathways; for example:
rosiglitazone COEXISTS_WITH ppar_gamma (Egerod et al., 2009; Johnson et al., 2000; Otake et al., 2011; Risérus et al., 2005) ASSOCIATED_WITH myocardial_infarction (Fliegner et al., 2008)
rosiglitazone INTERACTS_WITH ppar_gamma (Gao et al., 2007; Kim, 2006; Lee, 2003; Tsukahara, 2006; Tzameli, 2004) ASSOCIATED_WITH myocardial_infarction
rosiglitazone INTERACTS_WITH glycerol-3-phosphate_dehydrogenase (Suzuki et al., 2013) COEXISTS_WITH ppar.gamma (Muhlhausler, Duffield, & McMillen, 2007) ASSOCIATED_WITH myocardial_infarction
rosiglitazone INTERACTS_WITH glycerol-3-phosphate_dehydrogenase INTERACTS_WITH ppar.gamma (Ding, Nagai, & Woo, 2003) ASSOCIATED_WITH myocardial_infarction
rosiglitazone INTERACTS_WITH resistin (Jung et al., 2005) INTERACTS_WITH ppar.gamma (Patel et al., 2003) ASSOCIATED_WITH myocardial_infarction
Figure 4-9: The predications retrieved by reasoning pathway for rosiglitazone causing myocardial infarction with specifying semantic groups for concepts.
There were 2,618 distinct predication pathways about “triglycerides”, “LDL-C” and “PPAR-gamma” specifying 247 unique middle terms. Drilling down, Figure 4-10 shows the connecting concepts between LDL-C and myocardial infarction that fall along the reasoning pathways employed by the PSI-double+triple model. In each reasoning pathway, the middle terms were ranked using inverse document frequency, to approximate the weighting used by the predictive model. For each predication in these pathways, the source sentences from the literature were retrieved. For example, the article “A comparison of lipid and glycemic effects of pioglitazone and rosiglitazone in patients with type 2 diabetes and dyslipidemia” (Boyle et al., 2002) explains that rosiglitazone increased triglycerides compared with pioglitazone and has different effect on plasma lipids which may contribute to heart disease. Figure 4-10 shows the middle terms retrieved to justify that rosiglitazone may cause myocardial infarction via LDL-C.

The capacity to retrieve and organize knowledge in this way suggests a new paradigm for information retrieval in which information supporting a hypothesis of interest is automatically aggregated and organized at the conceptual level. However, as the number of assertions in the literature far exceeds the number of documents, further research is needed to develop methods through which to prioritize these assertions, and present them in a manner conducive to human consumption.
Figure 4-10: Middle terms that were retrieved by PSI discovery patterns involving LDL-C
4.5 Discussion

This study evaluates the ability of scalable LBD methods based on distributional semantics to rank the plausibility of connections between drugs and potential ADRs. I find that both the RRI and PSI models are able to retrieve known side effects of drugs, but PSI performs this task better, as one would anticipate given the additional information beyond co-occurrence that it encodes. The PSI model can further provide the reasoning pathways that were used to link a drug to a predicted side effect. Consequently, relevant literature can be retrieved to support the predictions, and provided to experts for review. However further research is needed to develop approaches through which the assertions underlying the large numbers of reasoning pathways utilized by the model can be prioritized for expert review, as these are too numerous for exhaustive manual review. Ultimately, I aim to provide domain experts with essential evidence while preventing information over-load. Even though it is not the best performing model, the RRI model has the advantages of a simple training process and the availability of more data to draw upon (as MetaMap has higher recall for concepts than SemRep has for predications). Conversely, the PSI model has the advantage of modeling plausibility, a capability with the potential to assist expert clinical review for pharmacovigilance. In addition, the correlation analysis between groups suggests that RRI and PSI complement each other, and can potentially be combined to improve performance on this task.

For predicting ADRs, several statistical models and Machine Learning (ML) algorithms have been evaluated against an edition of SIDER, or a subset of this repository. In addition to methodological differences, these approaches have leveraged different data sets and a variety of knowledge bases as a basis for making predictions. In the section that follows, I
will provide a review of these approaches, and the performance they have documented for the prediction of ADRs in SIDER.

Pauwels et al. represented drugs using as features the presence or absence of chemical substructure components described in PubChem (B. Chen, Wild, & Guha, 2009). In addition to standard supervised ML approaches, they applied canonical correlation analysis (CCA), including a sparse variant that emphasizes a small number of informative features for each training example. These methods were used to predict SIDER side effects, with a reported global AUC of 0.8932 (Pauwels et al., 2011) on a set of 1350 ADR and 888 drugs, using fivefold cross-validation.

Subsequently, Liu et al. applied five supervised ML algorithms to the same SIDER set. In addition to the PubChem-derived chemical substructure features used by Pauwels et al., features were drawn from DrugBank (Knox et al., 2011) (drug targets, transporters, and enzymes), KEGG (Kanehisa & Goto, 2000) (pathway information) and SIDER itself (drug indications and side effects). A classifier was built for each SIDER ADR, and the classifiers were then evaluated on 832 SIDER drugs (for which DrugBank IDs could be found) using fivefold cross-validation. The support vector machine (SVM) algorithm performed best with a global AUC of 0.9524 on the full SIDER dataset (Liu, Wu, et al., 2012). The authors attribute much of the improvement in performance by this and other metrics to the effects of incorporating SIDER side effects as features, suggesting that certain side effects have a tendency to co-occur in drug label data.

Other authors have reported performance on subsets of SIDER using similar methods. For cardio-toxicity related ADRs in SIDER, a median AUC of 0.771 using SVM for prediction has been reported (Huang et al., 2011). In this case, features were selected from information
about intended drug targets in DrugBank, and information about off-target effects from an expanded protein-protein interaction network developed using gene ontology (GO) annotations. A SVM classifier was built for each evaluated ADR and cross-validated on SIDER.

With respect to performance, two of these studies, Pauwels et al. (Pauwels et al., 2011) and Liu et al. (Liu, Wu, et al., 2012) report a global AUC of close to 0.9 or higher with the best of their methods. Though the results are not directly comparable as I made predictions on a per-drug rather than a per-ADR basis, the difference between the global AUC of these methods and that obtained with our approach seems large. However this difference in global AUC is misleading. As noted by Liu and colleagues in their paper, the imbalance between positive and negative examples across ADRs and the way in which the global AUC was calculated in this work leads to an apparent inconsistency between it and the other evaluation metrics presented. For example, Pauwels and his colleagues display the AUC across different ADRs in a series of box plots, which shows a median AUC for the best-performing method (by this metric) of slightly above 0.6. Acknowledging this issue, Liu et al. also report precision and recall for each evaluated method with, for example, precision of 0.66 and recall of 0.63 for SVMs with their maximal feature set. Notably, the AUC in this case was around 0.95.

This apparent inconsistency can be explained by the effect of the prevalence of positive examples for each ADR on the prediction strength. This is readily apparent for simple algorithms such as Naive Bayes, where the prior probability of a given category is incorporated into the estimate. However, it is also an issue for more sophisticated algorithms such as SVM (Akbani et al., 2004) particularly when the imbalance between
categories is severe. This is the case for many of the ADR examples: Liu et al. report a positive to negative ratio of around 1:166 for 554 of the 1135 ADRs. So given the same set of features, instances in these cases are likely to receive a lower prediction score than those in balanced cases. When these scores are aggregated across examples to generate a global AUC, ML methods that incorporate the category bias will obtain an inflated global AUC on account of this tendency to assign lower scores to instances with few positive examples. However, as noted by Liu et al. and demonstrated by the other reported metrics, this AUC is not an accurate reflection of the ability of these models to detect positive examples.

To simulate the effects of category bias on global AUC, I performed a simple experiment in which I multiplied the similarity scores produced by our model by the proportion of positive examples for each drug. This roughly approximates the effects of an accurately estimated prior class probability during cross-validation experiments. This resulted in an increase of our global AUC from 0.68 to 0.88. I do not present this result for the purpose of comparative evaluation, as our experiments are not directly comparable with prior ML work for other reasons I will subsequently discuss. Rather, I present it as an illustration of the disproportionate influence of category bias on global AUC, which underscores the issues with this evaluation metric raised by Liu et al. I trust it will also serve to dispel the misleading impression that the predictive accuracy of our methods is vastly inferior to that reported previously.

As our method does not consider the number of ADRs associated with a particular drug, the global AUC and median AUC approximately agree with one another. Our median AUC (across all drugs) of 0.7058, which falls somewhere in between that reported by Pauwels et al. (Pauwels et al., 2011) (across all ADRs) and Huang et al. (Huang et al., 2011) (across
cardio-toxicity related ADRs only). On account of the difference in denominator these results are not directly comparable, but they do further illustrate the discrepancy between global and local AUC in models that are not agnostic to class imbalance. Arguably such agnosticism is desirable from the perspective of an expert review, as it is difficult to justify the assertion that those drugs with fewer known associated side effects should be considered less likely to cause some newly observed side effect (and vice versa).

With respect to methodological differences, all of the above methods are supervised ML methods, and were applied to infer whether or not drugs were associated with each ADR from the features of other drugs known to be associated with this ADR. So the predictive models were generally customized on a per-ADR basis, for example by generating an individual classifier for each ADR in the case of SVM. In contrast, our approach infers a set of abstract reasoning pathways that were consistent across the drugs I evaluated. However, as illustrated by the absence of evidence across certain pathways in the rosiglitazone example, some pathways may be more predictive for particular medications or ADRs. So it seems likely that I could further improve our performance by incorporating supervised ML, a direction I plan to explore in future work.

Our approach differs with respect to the knowledge sources utilized also. For example, KEGG and DrugBank are manually curated databases. Our knowledge base, SemMedDB, contains predications that have been automatically extracted by SemRep from the biomedical literature using NLP. Inaccuracies in language processing, or indeed in the literature itself may introduce sources of error that are not present in manually curated data. However, the scope of the literature is much broader than that of human-curated resources. Furthermore, as there is no agreed-upon gold standard for ADRs, different studies have
utilized different datasets as reference sets (Tatonetti et al., 2012). Our study employed SIDER2, which includes considerably more drugs and ADRs than SIDER1.

This work has several limitations. The first of these concerns the use of SIDER2 as a reference standard. As SIDER2 consists of recognized side effects only, I cannot reliably distinguish between false positive signals and previously unknown ADRs. Furthermore, SIDER was compiled from package insert information by NLP tools (Kuhn et al., 2010), and as such may include side effects that seldom occur in practice or false associations that were caused by text-mining errors (Kuhn et al., 2013). While SIDER2 is sufficient to evaluate the hypotheses of the current work, in future work I plan to incorporate other data sources, such as EHR data and FDA reports. These data sources may provide additional evidence to support the assertion that an unknown drug/ADR pair is worth investigating further. Alternatively, they may provide the means to select a subset of the side effects in SIDER2 that have been observed frequently in practice as an additional evaluation set.

Secondly, the MMB repository contains one year less literature than the SemMedDB dataset. There is a difference of 1,757,64 citations (7.9% of SemMedDB dataset). These were the newest datasets at the time of the experiment. However the MMB repository has many more data points than SemMedDB. For example, more than 99.99% of citations have concepts extracted by MetaMap and 59.91% of citations have predications extracted by SemRep.

Another concern is the existing knowledge about causal relationships between drugs and related ADRs from the literature. For our dataset (90,787 pairs), 45% of pairs (concerning 953 drugs) co-occur directly in the MMB repository and 5% of pairs (concerning 693 drugs) have direct causal relationship (drug CAUSES ADR) in SemMedDB. So PSI’s
accuracy is dependent upon its ability to meaningfully infer connections between concepts that were not previously linked in its database, a capacity that would be particularly useful as a means of assessing novel ADRs that had not previously been documented in the literature. RRI is also able to draw such inferences, but in this case more of its performance may be attributable to direct co-occurrence.

Inspecting the middle terms that our model retrieved for rosiglitazone-myocardial infarction association (Figure 4-10), I found that at times uninformative high-level concepts, such as “genes” and “proteins”, were retrieved. In our study, I addressed the issue of uninformative high level concepts in two ways, both related to their propensity to occur relatively frequently in the corpus. Firstly I used a frequency threshold of 1,000,000 to exclude frequently occurring concepts contained in SemMedDB. The frequency of “genes” and “proteins” is less than the threshold and cannot be filtered. Secondly, I used a weighting procedure to reduce the influence of high-frequency terms on the training process. However, more sophisticated approaches to filtering are possible. Information concerning UMLS semantic types and position in the UMLS hierarchy could be used to develop more sophisticated approaches, to further filter out uninformative high-level concepts, which may improve performance.

The predictions made by PSI depend upon assertions extracted from the biomedical literature. One concern about the extracted predications is that they may be implausible on account of NLP errors. Though SemRep has been optimized for precision, its precision is not perfect. For example, Kilicoglu and colleagues estimate the precision of SemRep to be around 0.77 (Kilicoglu et al., 2008). Based on this, and other published evaluations (Ahlers, Fiszman, Demner-Fushman, Lang, & Rindflesch, 2007; Kilicoglu et al., 2010), it is
reasonable to estimate that around three in four predications in the set are perfectly accurate. In many cases, inaccurate predications nonetheless indicate co-occurrence, which is also informative. The PSI-based analogical reasoning approach I have employed is robust to isolated language processing errors, as highly ranked predictions are based on assertions extracted from thousands of unique reasoning pathways. For example, for the rosiglitazone-myocardial infarction association, 108,100 unique predications were retrieved, spanning eight of the inferred reasoning pathways. On average, individual predications were supported by 17 excerpts from the literature. If I extrapolate from prior published evaluations of SemRep, the predication concerned would have been accurately extracted from around 12 of these excerpts. So it is likely that at least some of the evidence supporting each individual assertion is accurate. Moreover, as this method is distributional in nature, it does not require that these assertions be perfectly accurate. Rather, the frequency with which an assertion is extracted factors into the strength of its contribution to a reasoning pathway. Nonetheless, the biomedical literature may contain controversial assertions, or contradictory conclusions from different experts or different experiments. This is illustrated by the rosiglitazone (brand name: Avandia) case. In 2007, the FDA added a black-box warning for heart-related risks to Avandia based on a meta-analysis (Nissen & Wolski, 2007) and three other studies (U.S. Food and Drug Administration, 2007). In 2013, the FDA lifted certain Avandia prescribing restrictions based on the readjudicated results of the RECORD trial (R. D. Lopes et al., 2013; Mahaffey et al., 2013), claiming the initial concerns were overblown (U.S. Food and Drug Administration, 2013). This decision was condemned by one of the authors of the original meta-analysis (Thompson, 2013). Currently our models weight the contribution of assertions using statistics related to local
and global frequency. However, it would also be possible to weight the importance of these assertions based on some assessment of the reliability of the source. For example, in information retrieval experiments, an approach incorporating citation information was better able to identify articles considered as important in a pre-existing bibliography (Herskovic & Bernstam, 2005). Possibilities include weights derived from the citation count of the source article, the impact factor of the journal, or the nature of the experiment described. It is possible that weighting metrics of this source would improve the predictions of our models, and they also suggest approaches to prioritize the large numbers of assertions supporting our predictions for review by human experts.

4.6 Conclusion
In this research, an emerging, scalable method of LBD that uses distributional statistics to infer and apply discovery patterns was adapted to evaluate the plausibility of drug/ADR relationships for the purpose of pharmacovigilance. The effective application of large amounts of partially accurate biomedical knowledge to this problem was facilitated by the scalable and robust nature of approximate inference in geometric space. This approach was shown to be more effective than a comparable co-occurrence based baseline, and has the further benefit of permitting the retrieval of evidence underlying the assertions used by the system to make its predictions. Consequently, our approach provides the means to assist with expert clinical review by providing evidence supporting the plausibility of the connection between drugs and ADRs. Furthermore, the models I have developed can be applied to filter drug/ADR signals that are detected in spontaneous reporting systems or EHR data, a direction I plan to explore in future work.
Chapter 5: Toward a Reality-based Repository of Adverse Drug Reactions:  
Comparing SIDER2 with Side Events Encountered in Practice

Recent informatics research has focused on the development and evaluation of automated approaches to pharmacovigilance (D. J. Almenoff et al., 2005; Andrew Bate, 2003; Wang et al., 2009). However, since the purpose of pharmacovigilance is to monitor and predict previously unknown side effects, there is no complete reference set with which to validate detected drug/ADR associations. Nonetheless, there is a need for such a gold standard in pharmacovigilance (Manfred Hauben & Aronson, 2007). For example, a reference set is required to evaluate a signal detection algorithm or validate the predictions of a pharmacovigilance system. As it is not practical to construct a large-scale validation set consisting of previously unknown side effects, common strategies involve validation against published drug reference book (Lindquist et al., 2000), drug safety related labelling changes (Manfred Hauben & Reich, 2004), curated reference sets (Coloma et al., 2013; LePendu et al., 2013; Ryan et al., 2012; Ryan, Schuemie, Welebob, et al., 2013), or large-scale database of known drug/ADR associations (Kuhn et al., 2010).

The EU-ADR (Coloma et al., 2013) and the Observational Medical Outcomes Partnership (OMOP) (Ryan et al., 2012; Ryan, Schuemie, Welebob, et al., 2013), major drug surveillance projects in Europe and America, systematically developed small curated reference sets for evaluating their new developed methods on drug safety surveillance with observational healthcare data (for example, administrative claims and EHRs). The EU-
ADR reference standard (Coloma et al., 2013) contains 94 drug-event associations (44 positives and 50 negatives). This was constructed based on literature, WHO Vigibase™ (WHO spontaneous reporting system (SRS)) and clinical adjudication. These associations were restricted to ten important events in pharmacovigilance and are adequately represented in the EU-ADR network. With the experiences of previous OMOP experiments and stakeholder interest, OMOP selected four events to construct a reference standard for methodology evaluation. The OMOP reference standard (Ryan, Schuemie, Welebob, et al., 2013) contains 399 drug-event associations (165 positive controls and 234 negative controls) that were developed from drug labeling, Tisdale level of causative evidence review (James E. Tisdale & Douglas A Miller, 2010) and systematic literature review. These associations are also required to have sufficient representations in OMOP databases. These two reference standards have been applied to evaluate different statistical methods in discriminating true effects from false drug-event associations with the EU-ADR databases (Schuemie et al., 2012, 2013) and the OMOP databases (William DuMouchel, Ryan, Schuemie, & Madigan, 2013; Madigan, Schuemie, & Ryan, 2013; Norén et al., 2013; Ryan, Schuemie, Gruber, Zorych, & Madigan, 2013; Ryan, Schuemie, & Madigan, 2013; Suchard et al., 2013).

In addition to the curated relative small reference sets, researchers have used the Side Effect Resource (SIDER) to serve as a gold standard in a large scale to validate detected ADRs or to evaluate drug-event signal detection systems (Deftereos et al., 2011; Saiakhov, Chakravarti, & Klopman, 2013; Yates & Goharian, 2013). SIDER contains information about the side effects of drugs that has been extracted from package inserts by text mining tools. However, there is reason to believe that SIDER may include false drug/ADR
associations, which would lead to inaccurate evaluation of the systems concerned. It has been argued that package inserts over-report with respect to side effects (Duke J et al., 2011). Analysis of SIDER has shown that text mining errors may occur when processing the package insert; for instance, generic warnings have been mistakenly extracted from labeling information (Kuhn et al., 2013).

Therefore, to use a large scale reference standard in pharmacovigilance evaluation, it is of interest to determine the extent to which the side effects reported in SIDER occur in practice. By limiting to the side effects that have sufficient representations in practice, false drug/ADR associations can be eliminated from SIDER and the predictive power of pharmacovigilance methods can be less affected by inadequate sample size of drugs or events. The information of practice usage can be obtained from a post-marketing SRS. A SRS is designed to collect anecdotal case reports. Even though they are not peer-reviewed, the case reports in SRSs can provide supporting evidence for possible drug/ADR associations (J. K. Aronson & Hauben, 2006). These collected data represent suspected drug/ADR associations reported by healthcare practitioners that observed a reaction in a patient under their care, by pharmaceutical companies that are mandatory required to report any collected side effects, or by consumers that may experience unpleasant reactions for drug treatments. Therefore, these reports provide anecdotal evidence that a side effect mentioned in a repository has occurred in practice.

In this study, I evaluate the frequency with which SIDER2 (SIDER version 2) drug/ADR associations occur in reports submitted to the Food and Drug Agency (FDA) Adverse Event Reporting System (FAERS). The ultimate goal of this research is to develop a drug/ADR
reference set consisting of those label-derived side effects that have been observed in practice, for the purpose of pharmacovigilance research.

5.1 Materials

SIDER2 is a publicly available database containing information on marketed medicines and their known adverse reactions (Campillos, Kuhn, Gavin, Jensen, & Bork, 2008; Kuhn et al., 2010). The current version (SIDER2) was released in 2012, and was used for this study. The information in SIDER2 was extracted from drug labeling information (package inserts) using text mining tools and a side effects dictionary. The source of package inserts includes British Columbia Cancer Agency, Facts@FDA, FDA Center for Drug Evaluation and Research, FDA MedWatch, and Health Canada Drug Product Database (DPD). Labeling information is from clinical trials, post-marketing surveillance, etc. To construct the side effects dictionary (Kuhn et al., 2010), Coding Symbols for Thesaurus of Adverse Reaction Terms (COSTART) was used as a seed dictionary and then was expanded by extracting synonyms for the seed dictionary from the Unified Medical Language System (UMLS). To build SIDER drug-event associations, the drugs’ indications and side effects related sections were processed to extract terms that correspond to the side effects dictionary and the terms extracted from indications were subsequently excluded from drugs’ side effects. SIDER2 contains 99,423 drug-event associations for 996 drugs and 4192 side effects.

FAERS is a spontaneous reporting system database that contains information on adverse event and medication error reports that were submitted to the FDA by pharmaceutical companies, health professionals and consumers in the United States (Center for Drug

5.2 Methods

5.2.1 Import SIDER2 and FAERS data in database

SIDER2 was downloaded and imported into a local database. Distinct drugs and side effects were retrieved from the “meddraAvderseEffects” table. Publicly available FAERS quarterly data files (2004 to 2012) were downloaded from the FDA website (Center for Drug Evaluation and Research & U.S. Food and Drug Administration, 2014). I imported all data into a SQL database aside from the 2012q4 data, since the metadata used in this quarter were inconsistent with the database as a whole. The database contain 4,070,077 FAERS reports. For this study, I used information in the “Drug” and “Reaction” tables.

5.2.2 Parsing drug and side effect terms using MetaMap (Figure 5-1)

2013 edition of MetaMap (MetaMap 2013 was used to process and annotate retrieved drugs and reactions and map them to UMLS concepts (U. S. National Library of Medicine, 2009) for the purpose of mapping between FAERS and SIDER2. MetaMap identifies matches between terms in text and candidate concepts from the UMLS. Each of these candidates is annotated with a confidence score, a Concept Unique Identifier (CUI), a preferred label, and one or more semantic types. The semantic types indicate the type of the concept.
Examples include “Pharmacologic Substance” or “Diagnostic Procedure”. Candidates can be assigned multiple semantic types.

I manually reviewed distinct semantic types or semantic type combinations for MetaMap processed SIDER2 drugs and side effects and found out two features of drugs or side effects relevant semantic types. First, there are semantic types that are irrelevant to my research, such as “Organ or Tissue Function”, “Organism”, or “Activity”. Candidates of those semantic types were excluded. Second, a combination of semantic types could often contribute a more relevant classification. For example, the semantic type “Immunologic Factor” occurred very frequently and often was associated with concepts that were not relevant to drugs. However “Immunologic Factor” in combination with “Pharmacologic Substance” provided more insight, and identified a relevant candidate in the context of our study. Another example of semantic type combinations is a concept annotation that has a candidate with only one semantic type having a higher confidence score and a candidate with semantic type combinations having a less confidence score. The manual review revealed that candidates with combinations of certain semantic types yielded more relevant concepts. Consequently, restricting the result list by only accepting very specific semantic types or combinations may increase the relevance of the concepts identified by MetaMap to the study.
5.2.3 Mapping between SIDER2 and FAERS

After annotating SIDER2 and FAERS with selected semantic types or semantic type combinations, corresponding CUIs were retrieved and used to find the matched FAERS drugs or ADRs for parsed SIDER2 drugs or ADRs. Figure 5-2 illustrates this process using the example drug “dipyridamole”. For this drug, 78 FAERS drug inputs were found.
5.2.4 Retrieve number of reports for mapped SIDER2 drug-side effect pairs and perform disproportionality analysis to find significant SIDER2 drug/ADR associations

For each SIDER2 drug-side effect pair, I retrieved the number of related FAERS reports through their mapped FAERS drugs or reactions. Then I constructed a contingency table for each drug-side effect association with corresponding FAERS terms. Disproportionality measures were used to calculate the significance of the association between drug and side effect pairs co-occurring in the FAERS reports. These methods operate under the assumption that if a drug causes a side effect, this drug and this event are more likely to appear together than random combinations of drugs and suspected adverse events in the SRS reports (Cao et al., 2005).
Routinely used disproportionality measures for pharmacovigilance drug-side effect signal detection include the Reporting Odds Ratio (ROR), Proportional Reporting Ratio (PRR), Yule’s Q, and relative risk (RR) (Balakin & Ekins, 2009; Cao et al., 2007; Emmanuel Roux et al., 2005; van Puijenbroek et al., 2002). In addition to these measures, I applied Chi-square test with the False Discovery Rate (FDR), which was introduced to measure the multiple-hypothesis testing error and has been successfully used in large-scale genomic studies to control false positive results (Benjamini & Hochberg, 1995; J. D. Storey & Tibshirani, 2003; J. D. Storey, 2002).

5.2.5 Performance measure

Descriptive statistics of agreement for drug-side effect relationships between SIDER2 and FAERS were estimated. We also conducted a manual review of a random sample of 60 pairs that occurred in SIDER2 without accompanying FAERS reports, to investigate the cause of this mismatch. DailyMed was used to retrieve drug package inserts and review the “Adverse Drug Reaction” section to track down the side effect of interest. MEDLINE was queried to find published literature or case reports that relate to the drug/ADR pairs retrieved from the package inserts.

5.3 Experimental design

To identify the subset of SIDER2 drug/ADR pairs that have been reported in practice, I compare these pairs with anecdotal evidence from FAERS reports (Figure 5-3). To map SIDER2 drugs to text in the FAERS “drug” field, and SIDER2 side effect to text in the FAERS “reaction” field, MetaMap 2013 was used to process all terms and a manual review process of selecting related semantic types that (Figure 5-1) was used to select appropriate
UMLS candidates. Multiple disproportionality analysis methods (A. R. Aronson & Lang, 2010; A. R. Aronson, 2001; Center for Drug Evaluation and Research & U.S. Food and Drug Administration, 2012) were then applied to analyze the significance of co-occurring drug-side effect associations. The significantly reported SIDER2 drug/ADR associations that are detected by all disproportionality measures are considered as a pharmacovigilance reference standard.

**Figure 5-3:** Study design for decreasing false SIDER2 drug-side effect associations

### 5.4 Results

#### 5.4.1 Concept extraction

Using the mapping procedure, I mapped 99% (986/996) of SIDER2 drugs and 97% (4072/4192) of SIDER2 side effects to UMLS concepts; 78% of FAERS drug strings (259,806 drugs) and 98% of FAERS reaction strings (15,989 reactions) to UMLS concepts. Among SIDER2 mapped UMLS concepts, 969 drugs and 3853 side effects were mapped to FAERS drug and reaction inputs. These concepts constituted 94.85% (94,306) of the SIDER2 drug/ADR pairs, which were subsequently compared with FAERS reports. Of
these 94,306 SIDER2 drug-side effect pairs, 11,306 pairs do not co-occur in FAERS reports. 83,000 drug/ADR pairs have co-occurring reports in FAERS database.

For this study I only included reports that were related to these UMLS concepts. Consequently, 4,070,225 reports were used for disproportionality analysis.

5.4.2 Reported SIDER side effects

I analyzed the frequency distribution of reports related to 83,000 SIDER2 drug/ADR pairs that have co-occurred in FAERS reports (Figure 5-4). The number of FAERS reports ranges from 1 to 360,146, with mean of 1944, median of 85, 1st quartile at 14, and 3rd quartile at 598. The histogram of representing all pairs has long right tail and data are skewed. For a more granular picture of this distribution, I plotted the histogram for each quartile (Figure 5-4). Most drug/ADR pairs have a relatively small number of FAERS reports and less than 600 FAERS reports.
5.4.3 Statistically significant drug/ADR pairs detected by disproportionality measures

I evaluated the significantly reported instances of SIDER2 drug/ADR associations by applying disproportionality measures using FAERS data (Table 5-1). About 60% to 80% of SIDER2 drug/ADR pairs met the thresholds for significance of these statistical measures. 46,203 SIDER2 pairs (904 drugs, 2984 side effects) were detected as statistically significant by all statistical metrics.
Table 5-1: The percentage of statistically significant SIDER2 drug/ADR pairs using disproportionality measures

<table>
<thead>
<tr>
<th></th>
<th>ROR-1.96SE&gt;1</th>
<th>PRR-1.96SE&gt;1</th>
<th>YulesQ-1.96SE&gt;0</th>
<th>IC-2SD&gt;0</th>
<th>Chi-Square test with FDR (q Value&lt;0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>60% (50,139/83,000)</td>
<td>60.9% (50,570/83,000)</td>
<td>61.4% (50,978/83,000)</td>
<td>55.7% (46,209/83,000)</td>
<td>79.7% (66,178/83,000)</td>
</tr>
</tbody>
</table>

5.4.4 Manual review sample drug/ADR pairs without case reports

I randomly selected 60 drug/ADR pairs from 11,306 pairs that were not included in any FAERS report and analyzed possible reasons for why there are no case reports (Figure 5-5). Only 25 pairs were listed as side effects in the drug labeling or package inserts. Among those 25, eight side effects were listed as “infrequent side effects” or less than 1% of tested patients. For three side effects, the trial provided limited evidence whether they were caused by the drug under trial. In two instances (cidofovir related hypophosphatemia and glycosuria), participants were on several other medications and so it is difficult to establish the causal relationships. In another instance (teniposide related arrhythmia), there was only one report of this complication during pre-marketing trials, it was presumed rather than confirmed clinically, and it had occurred in an elderly patient with a variety of other health problems. Another six pairs among the 25 pairs had either mild side effects (e.g. dry throat, dizziness), or concerned a rarely prescribed drug (e.g. maraviroc). I could not find a plausible explanation for the remaining eight side effects that did not occur in any FAERS report; these side effects were listed in “Adverse Reactions” section for different systems without frequency information in the package inserts.
Among the remaining 35 drug/ADR pairs, six did not concern drugs; they instead concerned a vitamin or some other chemical component. So no package insert was available specifically the six chemical components (e.g. 1,25-dihydroxyvitamin D₃, betaine). For seventeen pairs package inserts were available, but the labeling does not list the side effects concerned. Four pairs suggested a possible text mining error, as SIDER2 side effect was close to the information in the labelling either conceptually or orthographically, but is, in fact, not a match. For four pairs that are drugs (e.g. ofloxacin), I couldn't find the package insert in the DailyMed database.

Figure 5-5: Interpretation of 60 randomly selected SIDER2 pairs without any supporting reports

5.5 Discussion

The study suggests that a higher precision subset of SIDER2 pairs could be identified by filtering out possible false positives using an inclusion criterion based on statistically significant association in the FAERS data. These false positives may result from NLP
errors during the parsing of drug labeling or inserts; on account of the inclusion of infrequent side effects that were not conclusively linked to the drug, but nonetheless occurred during clinical trials; or for other reasons. For the evaluation of new methods in pharmacovigilance it is imperative to use a high quality data set. If a data set used as a gold standard contains uncertainties itself the performance assessment of a new system or algorithm against these data set is difficult. It cannot be clearly concluded as to why the results of a new method show a certain outcome. Both a bias towards good or bad results could be caused by the quality of the reference data rather than by the method itself. Furthermore, in the case of machine learning systems trained on these data, the inaccuracies inherent in the data set may impair the performance of the derived models.

There are some limitations to this study. With the mapping procedure, it frequently occurred that input strings, specifically those of reactions, were mapped to multiple concepts. This was the case when either no concept existed that expressed the entirety of the input, or MetaMap simply could not recognize it. For example, “carcinoma of the small bowel” exists as a concept in and of itself, even though it is a combination of carcinoma as a clinical observation and small bowel as a body location. On the other hand, for “splenic embolism” MetaMap lists a candidate concept for each of embolism and spleen. The spleen, which is of semantic type “Body Part, Organ, or Organ Component” is certainly not an appropriate result for splenic embolism, while the embolism itself is too general as it may appear at other body sites also. So rather than accepting “embolism” by itself, a combination of those two concepts would be most appropriate. The main concept, embolism in the example, and one or more concepts that contribute, modify, or tag the main concept could be combined. Similarly, for the input string “neutrophil count
increased” MetaMap detects “neutrophil count” as “Laboratory Procedure” and “increased” as “Quantitative Concept” but without connection to the neutrophil count. Considering a combinations of concepts of different semantic types that MetaMap treats as atomic units may improve the quality of our results.

It also has been argued that SRSs are vulnerable to bias, and that the ADR under-reporting rate is considerable. Heterogeneity and bias result from different reasons: severe effects were more reported than mild ones, and labeled adverse drug reactions are more likely to be reported than unlabeled ones (Benjamini & Hochberg, 1995). The first of these findings is supported by our interpretations of 60 randomly selected SIDER2 drug/ADR pairs examined without FAERS reports, as some of these side effects were mild in nature. That labeled side effects are more frequently reported supports our argument for the use of FAERS reports as a means to validate drug/ADR repositories.

An important outcome of this study is a high-precision subset of SIDER2 in which reported drug/ADR pairs are supported by the statistical significance of their FAERS reports. I anticipate that this subset will be of value for the training and evaluation future pharmacovigilance systems and new ADR signal detection algorithms.

5.6 Conclusion

Based on the consistency between FAERS and SIDER2 drug/ADR associations, a drug/ADR reference set is retrieved. The utility and application of the reference set will be further evaluated in pharmacovigilance research.
Chapter 6: Improving Signal Detection from the Electronic Health Records by

Using Literature Based Discovery

In previous experiments, I demonstrated that (1) possible drug/ADR signals can be detected from EHR data; and (2) LBD models based on methods of distributional semantics can identify plausible ADRs. In this experiment, I test my unifying hypothesis that signal detected from EHR data can be improved by integrating knowledge from the literature. Specifically, I propose that the performance of signal detection from the EHR can be improved by reranking the detected signals using LBD models based on distributional semantics. I evaluate this procedure using the drug/ADR reference set that we developed from SIDER2 and FAERS, as well as with SIDER2 itself.

6.1 Experimental design (Figure 6-1)

In Chapter 3, I described how Drug/ADR signals were detected from EHR data hosted in a Clinical Data Warehouse (CDW) using several disproportionality measures and other statistical tests. The best performing statistical model from those experiments was chosen for this one. Likewise, the best performing LBD model from the experiments described in Chapter 4 was utilized for this experiment. To perform the experiment, drug/ADR signals were detected in EHR data using the statistical model and ranked in accordance with the strength of their associations (according to the statistical model). These drug/ADR signals are referred to as pre-reranking drug/ADR signals. Subsequently, these signals were
reranked in accordance with the similarity scores (the measure for plausibility) that the LBD distributional model estimated from the relatedness between drugs and ADRs across the set of discovery patterns identified in the experiments described in Chapter 4. Consequently, these are referred to as post-reranking drug/ADR signals. They were then evaluated by comparing the pre- and post-drug/ADR signals to both SIDER2 and a reference set that we developed from it.

Figure 6-1: Overall research design for reranking statistically significant drug/ADR associations by similarity scores

6.2 Materials and Methods

6.2.1 Models selected for this study

I have elaborated methodologies for statistical data mining from EHR and LBD distributional semantics in Chapter 3 and Chapter 4, respectively. Of these methods, the Chi-square statistics with FDR (for estimating the statistical significance of observed
associations) and PSI double+triple group were best performing models in their respective experiments, and consequently were utilized for this one.

In pre-reranking drug/ADR signals, the chi-square score is used as to reflect the statistical strength of the association observed in CDW data. The FDR-based q values are used as a significance threshold, to predict if the signal is true or false. Similarity scores from PSI double+triple model are used to rerank those signals that fell above the q-value threshold.

6.2.2 Experimental reference standards and test dataset

I evaluated the above models using two reference standards – side effect resource (SIDER2) and the reference set I developed using SIDER2 and FDA adverse event reporting system (FAERS, as described in Chapter 5). The reference set is relatively small and contains only those SIDER2 relationships that frequently occurred in SRS databases. For each reference standard, its drugs and side effects that are not only contained in the CDW EHR data but also represented in the PSI model were eligible for this experiment. This resulted in a set of 811 drugs and 1879 ADRs with SIDER2 and 773 drugs and 1374 ADRs with the reference set (Table 6-1). Among these, only the predicted positive pairs by Chi-square statistics with FDR as statistically significant associations were utilized for the analysis.

Table 6-1: SIDER2 and a Reference Set are used to construct the dataset for the experiment

<table>
<thead>
<tr>
<th>Group</th>
<th>Drugs</th>
<th>ADRs</th>
<th>True Pairs</th>
<th>False Pairs</th>
<th>Total Pairs</th>
<th>Statistically significant associations</th>
<th>True positives in statistically significant associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIDER2</td>
<td>811</td>
<td>1879</td>
<td>260,555</td>
<td>395,468</td>
<td>436,865</td>
<td>260,555</td>
<td>28,114</td>
</tr>
<tr>
<td>Group</td>
<td>Drugs</td>
<td>ADRs</td>
<td>True Pairs</td>
<td>False Pairs</td>
<td>Total Pairs</td>
<td>Statistically significant associations</td>
<td>True positives in statistically significant associations</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------</td>
<td>------</td>
<td>------------</td>
<td>-------------</td>
<td>-------------</td>
<td>----------------------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>Reference Set</td>
<td>773</td>
<td>1374</td>
<td>40,838</td>
<td>315,544</td>
<td>356,382</td>
<td>215,024</td>
<td>27,765</td>
</tr>
</tbody>
</table>

### 6.2.3 Performance metrics

Average precision (AP) is the average of the precision that is measured at the rank at which each correct prediction is retrieved. Precision at 100 (Manning et al., 2008) is the precision at the first 100 retrieved results. A true positive drug/ADR association, defined as “rediscovery”, is an adverse effect that is confirmed by SDIER2 or the reference set. The median rank of rediscoveries across statistically significant drug/ADR associations approximates the point in the ranked list that half of the known adverse effects were recovered by Chi-square statistics with FDR or PSI-double+triple model.

AP, precision at 100, and median rank of rediscoveries are calculated and compared between pre- and post-reranking for the two datasets. A two-sample Kolmogorov-Smirnov test was conducted to evaluate if there is a significant difference for the total number of rediscoveries at each rank between pre- and post-reranking signals. The rediscovery at each rank is also plotted for different dataset.

To measure the performance with respect to the true positive rate (TPR) and false positive rate (FPR), receiver operating characteristic (ROC) curve was plotted for statistically significant drug/ADR associations. Subsequently, an area under the ROC curve (AUROC)
from the ranking of each model and an area for precision and recall (AUC for Precision-Recall) were calculated using AUCCalculator (Davis & Goadrich, 2006).

6.3 Results

6.3.1 Performance

Results of comparing pre-reranking and post-reranking for different datasets are shown in Table 6-2. PSI-based models perform better than RRI-based models and both models perform better than the random baseline.

Table 6-2: Results of precision and rank-based measures for different groups and different datasets

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Group</th>
<th>MAP</th>
<th>Precision at 100</th>
<th>Median Rank</th>
<th>AUC for Precision-Recall</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIDER2</td>
<td>Pre-reranking (Chi)</td>
<td>0.1424</td>
<td>0.16</td>
<td>103,528</td>
<td>0.1411</td>
<td>0.5708</td>
</tr>
<tr>
<td></td>
<td>Post-reranking (Similarity)</td>
<td>0.1640</td>
<td>0.21</td>
<td>85,166</td>
<td>0.1639</td>
<td>0.6323</td>
</tr>
<tr>
<td>Reference</td>
<td>Pre-reranking (Chi)</td>
<td>0.1655</td>
<td>0.19</td>
<td>87,218</td>
<td>0.1641</td>
<td>0.5664</td>
</tr>
<tr>
<td>Set</td>
<td>Post-reranking (Similarity)</td>
<td>0.1847</td>
<td>0.21</td>
<td>75,197</td>
<td>0.1846</td>
<td>0.6173</td>
</tr>
</tbody>
</table>
With the SIDER2 dataset, the median rank of true positives with EHR data (pre-reranking signals) is 103,528; and in combination with the LBD model (post-reranking signals) it is 85,166. With the reference set, the median rank of true positives in pre- and post-reranking signals is 87,218 and 75,197, respectively. Two-sample Kolmogorov-Smirnov test shows there is a significant difference of the accumulated number of true positives at each rank (Figure 6-2) between pre- and post- reranking methods for both datasets (all $P$-value < 2.2e-16).

![Graph of accumulated number of true positives](image1)

**Figure 6-2:** Comparing accumulated number of true positives for each ranking between pre- and post-reranking signals for SIDER2 set (top) and the Reference set (bottom)
6.3.2 AUC

Figures 6-3 and 6-4 present the global ROC curves. ROC curve shows the tradeoff between sensitivity and specificity. The AUC provides a cumulative estimate of accuracy, and is shown for each model in Table 6-2. With respective dataset, AUROC of post-reranking method is greater than the pre-reranking method.

Figure 6-3: ROC plot of true positive rate and false positive rate for pre- and post-reranking groups for SIDER2 set
6.4 Discussion

In this experiment, drug/ADR signals that were predicted from EHR data using the best performing statistical algorithm were reranked using the PSI model. This resulted in significant increases of the true positive rate at a corresponding rank (in comparison to the true positive rate with no reranking). Precision and AUC are better performed with post-reranking method. The experiment demonstrates that the PSI model can filter noisy signals using a measure of the plausibility of the relationship between the drugs and ADRs concerned.

Overall, based on what I have learned from the series of experiments, I want to propose the architecture for a plausibility-based pharmacovigilance system (Figure 6-5). This
framework describes the essential steps: (1) Process EHR data to extract coded drugs and possible problems; a NLP tool is required for processing unstructured clinical notes. (2) Select a cohort for drug/ADR analysis. (3) From the drug/problem candidates, existing known relationships knowledge is used to filter the known drug/problem relationships. (4) Conducting statistical analysis for drug/problem candidates and identifying statistically significant drug/ADR associations. (5) Using the PSI model to justify plausible drug/ADR associations and filtering the statistically detected signals using their plausibility. (6) Present plausible drug/ADR signals and their evidence retrieved by the PSI model to clinicians or practitioners for review. (7) Automatically submit the detected drug/ADR signals with all related clinical information to PV health administrative departments. This helps clinicians in making clinical decision for the patients that are taking the relevant drugs.

6.5 Conclusion

This experiment supports my overall hypothesis that the precision of signal detection from EHR data can be improved by integrating knowledge from the biomedical literature.
Figure 6-5: The proposed architecture for a plausibility-based pharmacovigilance system
Chapter 7: Key Findings, Innovation, Contributions, Future work and Conclusions

7.1 Overview and summary of key findings

For this thesis I conducted four experiments. First, of all the SIDER2 ADRs that occur in an outpatient EHR system, I was able to detect 10-15% in that EHR data by using existing disproportionality measures and Chi-square statistics with FDR. The Chi-square statistics with FDR was demonstrated as the best performing model with an F-measure of 0.1826, evaluated against SIDER2. Second, I built two LBD models based on scalable methods of distributional semantics (RRI and PSI discovery patterns) to identify possible drug/ADR associations utilizing the biomedical literature. The PSI discovery patterns model outperforms the RRI co-occurrence based model and can be used to evaluate the plausibility of drug/ADR associations. It has the additional advantage of modeling the relations between the involved medical concepts. Third, as a consequence of possible false associations that exist in the SIDER side effects dataset used for evaluation, I constructed a drug/ADR reference set. This reference set is based on the consistency between FAERS and the SIDER data set. I then used this as an additional reference set for evaluating side effects. Fourth and last, I applied a plausibility measure obtained through PSI discovery patterns to rerank the statistically significant drug/ADR associations detected in EHR data, which improved the precision against SIDER2 and the newly developed reference set. This verified my overall hypothesis that using literature to justify plausibility can improve the quality of signal detection from EHR data.
7.2 Innovation

I have developed means to partially automate the signal evaluation process by integrating knowledge extracted from the biomedical literature with possible drug/ADR signals derived from EHR data using statistical methods. To do so, recent LBD methods using “discovery patterns” were adapted to the task of evaluating the plausibility of drug/ADR signals at scale. To the best of my knowledge, this is the first time knowledge from the biomedical literature has been integrated with EHR data for signal detection. Previous attempts at applying methods of literature based discovery to find possible mechanisms to explain drug/ADR associations were conducted using manually defined discovery patterns. My research describes a method to automatically define pharmacovigilance related discovery patterns from the literature, and applies those to find predication-based explanations in an automated way on a large scale.

In summary, this research is the first to integrate semantic predications into the signal detection process to provide evidence supporting the plausibility of the connection between drugs and ADRs. The automated evaluation of plausibility to support causality according to the meaning defined by the Bradford-Hill criteria, is a novel contribution to the field of pharmacovigilance and to signal detection in general.

7.3 Theoretical contribution

From a theoretical perspective, my work is motivated by the notion of abductive reasoning as described by American philosopher CS Peirce (Peirce, 1955). According to Peirce, abductive reasoning (Schvaneveldt & Cohen, 2010) is the process of seeking the best possible explanation for an observation. The observation is a given and it is the goal to find an explanation that is
sufficient to justify the presence of the observation. This is in contrary to deductive reasoning where from a given starting condition possible consequences are deduced. It is also important to note the difference between a sufficient and a necessary condition. A sufficient explanation serves to explain an observation but its presence is not necessary. Thus, the observation can still be explained by other conditions and one explanation does not explicitly exclude others.

Currently in PV, signals indicating possible causal associations are discovered within a large database and then manually evaluated. A problem with this approach is that mined associations can be relevant or irrelevant, or may even be negative associations. Exhaustive manual review of these potential signals is not feasible. In my research I abduce explanations in an automated way, or, in other words, propose and find logical explanations for the identified associations in an automated way. The theoretical value added is then to use the generated explanations to collectively assess a measure for the plausibility of an association. The theoretical contribution is constituted by not primarily focusing on finding explanatory hypotheses but by focusing on assessing the plausibility of an observation. Where abductive reasoning is concerned with finding truthful explanations for an observation, my framework adds to that by going a step further and also assessing the truthfulness of the observation itself.

7.4 Practical contribution

Practically, the detection of signals from EHR data has the potential to improve post-marketing drug surveillance in real world applications. The computational requirements of the employed algorithms and tools are sufficiently low to allow for delivery of results in real time. In this context, “real time” effectively means timely surveillance, and thus the early detection of ADRs that are potentially harmful to patients. In the concluding experiment I showed that augmenting statistical
associations with a plausibility measure enhances the identification of known ADRs, suggesting that this approach would also lead to more accurate identification of novel ADRs. Furthermore, delivery of the underlying explanatory hypotheses to domain experts has the potential to increase the efficiency of the critical clinical review process. This is because the number of associations that is mined from EHR data is very high, and these signals still have to be manually evaluated. Prioritizing signals automatically has the potential to speed up the review process. Noise can be separated from those signals supported by plausible evidence because of the evidence provided to researchers or practitioners. Moreover, the evidence provided is in a very concise format (at the predication level rather than the document level) that allows for the exploration of a large amount of evidence in an efficient manner.

With respect to informatics in general, the methods and procedures of integrating formal knowledge can be generalized and applied to other domains. With the rapid growth of use of EHR data for clinical research (Hripcsak & Albers, 2013), new findings can be learned from the EHR data. The PSI discovery patterns model can be adapted and used to provide the automated interpretation of the findings. A very interesting example is outbreak surveillance, for which a timely identification of plausible signals is essential. For example, an influenza outbreak can be identified from EHR data by mining abnormal lab results, symptoms and outpatient diagnoses. PSI discovery patterns can retrieve possible sources, mode of transmission and risk factors (CDC, 2006) by analyzing the plausible pathways between the virus and the disease/symptoms. This can assist field investigators’ work.

In summary, the detection of signals from EHR data and the subsequent evaluation using supporting evidence from literature with PSI discovery patterns in pharmacovigilance on a large scale in an automated way has not been done before and has shown good results. The methods and
procedures proposed and examined in this work are novel to the field of pharmacovigilance and applicable to other informatics areas.

7.5 Future work

In future work, I plan to improve my methods for the estimation of plausibility, and provide better ways for domain experts to explore the evidence that supports the explanatory hypotheses the methods generate. Although predications are a concise way to present supporting evidence gathered by PSI discovery patterns from the medical literature, the presentation to domain experts for review in a concise way still presents a challenge. Since evidence is retrieved based on predications which can naturally be built into a graph network, it is straightforward to examine if graph algorithms can be utilized to analyze the evidence network. This can result in prioritization or the identification of important biological factors, or even biological pathways. Subsequently, improved visualizations could highlight specific aspects like biological factors and pathways and could put an emphasis on important concepts through their connectedness for example. This would add to the efficacy of the manual review process.

The methods developed for my research have the potential to support a real-time PV system. This work has demonstrated that signal detection and signal evaluation can be done in a partially automated and therefore, potentially more timely manner. I will continue working in this direction with the aim to contribute to the development of an active drug surveillance system.

7.6 Conclusions

This thesis demonstrates that drug/ADR associations can be detected from unstructured outpatient clinical notes by using statistical mining algorithms. PSI discovery patterns can further improve
the precision of detected signals by leveraging knowledge from literature and modeling the plausibility of the identified associations. Consequently this work has extended the state of the art in EHR-based pharmacovigilance and contributed new ideas that pave the way for further studies with the potential to further enhance the field of pharmacovigilance and drug safety.
References


Collins, F. S. (2011). Reengineering translational science: the time is right. Science Translational Medicine, 3(90), 90cm17–90cm17.


141


Kuller Lewis, Alice Arnold, Russell Tracy, James Otvos, Greg Burke, Bruce Psaty, … Richard Kronmal. (2002). Nuclear Magnetic Resonance Spectroscopy of Lipoproteins and Risk of Coronary Heart Disease in the Cardiovascular Health Study. Arteriosclerosis, Thrombosis, and Vascular Biology, 22(7), 1175–1180. doi:10.1161/01.ATV.0000022015.97341.3A


doi:10.1016/j.ahj.2013.05.005


doi:10.1056/NEJMp078057


doi:10.1016/j.jbi.2003.11.003


doi:10.1109/TITB.2005.855566


