

國立臺灣大學管理學院資訊管理學研究所



碩士論文

Graduate Institute of Information Management

College of Management

National Taiwan University

Master Thesis

從台灣健保資料庫偵測藥物副作用：使用學習排序法

Detecting Drug Safety Signals from National Taiwan Health
Insurance Research Database: A Learning to Rank Approach

謝采璇

Tsai-Hsuan Hsieh

指導教授：魏志平 博士

Advisor: Chih-Ping Wei, Ph.D.

中華民國 103 年 07 月

July 2014

國立臺灣大學碩士學位論文 口試委員會審定書

從台灣健保資料庫偵測藥物副作用：使用學習排序法

Detecting Drug Safety Signals from National Taiwan Health Insurance Research Database: A Learning to Rank Approach

本論文係謝采璇君（學號 R01725025）在國立臺灣大學資訊管理學系、所完成之碩士學位論文，於民國 103 年 7 月 28 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

楊所生

盧信銘

蕭非元

魏志平

所長：

蔡益坤

誌謝



碩士生涯兩年一晃眼就過了，如今也到達了完成這份論文的時刻，雖然我已盡自己所能進行這份研究，但仍有許多不足的地方需要老師及同儕提點與費心。

此份論文以及碩士班兩年學業的完成，最感謝的人莫過於我的指導教授—魏志平教授。魏老師教學態度十分謹慎，邏輯清晰，非常有研究精神，對於求知鍥而不捨，對於論文的架構與研究方法，老師總是給予我們空間思考，也會適時幫助我們，釐清論點的邏輯，讓我們不至於在研究中迷失方向。另外，老師也時常關心我們的日常生活、生涯規劃以及生活趣事，與我們分享人生經驗，並引導我們積極、正向面對未來的人生。也特別感謝三位口試委員：盧信銘老師、蕭斐元老師以及楊錦生老師，在口試時給予許多寶貴的建議，使本論文更臻完備。

再來，我要感謝實驗室大家長虹鈞，為我們這群小朋友張羅各項事物，不時提供實驗室各項資源，讓我們能專心致力的研究。感謝陳連進博士給予本研究許多資源，並提供許多寶貴的意見，讓這篇論文有好的開頭。感謝魏門的各位好夥伴：感謝光昇一起在實驗室研究各種跟論文相關的事務，一起在實驗室崩潰與成長；感謝黃葳不時關心我的論文與生活大事，並與我分享生活大小事；感謝蓓好時時督促大家的進度。感謝史達林、冠宇、尹安、鴻英、泰頤等學長姐給我的提點；感謝竣賢、宇婷、子勻、筑雅、小熊等學弟妹給我的幫助；感謝宗浩、傑凱、昭容與因篇幅關係無法逐一系列上的好朋友們給予我的鼓勵。最後，要壓軸感謝我的爸媽，對於我的決定都非常支持，也總是給予我最大的祝福，謝謝你們。

謝采璇 謹識

于台大資訊管理學研究所

西元二零一四年七月

中文摘要



藥物不良反應是一個相當嚴重的全球性醫療問題，造成許多病人必須再度就醫，更嚴重的情況甚至導致死亡。因此，近年來受到許多國家的重視，進而推動藥物上市後的監控。起初，藥物上市後的監控是利用藥物不良反應通報系統來進行分析，但藥物不良反應通報系統是一個由醫護人員或是病人自行通報的系統，因此資料品質較難維護，並且可能有個人偏誤的報告。而後藥物上市後的監控轉向分析電子健康紀錄(Electronic Health Records)，因為電子健康紀錄包含病患的電子病歷或健康保險申報相關之就診資料，其為持續性的資料，涵蓋的人口範圍也較廣，資料內容亦較準確。

台灣的健保資料庫是一個健康保險申報相關之資料庫，本研究以台灣健保資料庫來進行藥物副作用的分析與探測，希望能利用學習排序法將藥物和可能導致的不良反應(疾病)的關聯進行排序，找出可能的藥物不良反應。我們建立四個實驗情境以評估本研究所提出之方法效能，其結果顯示，本研究提出之方法能夠有效的提升探測藥物不良反應的準確度，可以提供給專業的醫藥學專家進行進一步的驗證與分析。

關鍵字：藥物不良反應、資料探勘、台灣健保資料庫

ABSTRACT



Pharmacovigilance (PhV) is a serious issue worldwide, because adverse drug effects are serious problems that cause harms to patients or even death. Traditionally, PhV research focuses on detecting adverse drug effects from spontaneous reports systems (SRS), which contains reports voluntarily reported by medical professionals, patients, and pharmaceutical companies. However, the volunteer nature of SRS databases causes some limitations (e.g., overreporting, data incompleteness). Thus, the PhV research starts to investigate the use of electronic health records (EHR) databases for drug safety signal detection in recent years. In this study, we propose a novel EHR-based drug safety signal detection method on the basis of the learning to rank approach. In addition to multiple disproportional analysis measures, our proposed method also incorporates as additional ranking variables that capture implicit relations between drugs and diseases for decreasing the importance of non-drug-outcome signals. We use Taiwan's national health insurance research database for drug safety signal detection. Our evaluation results suggest that our proposed method significantly outperforms existing disproportional analysis methods (each of which uses a single disproportional analysis measures).

Keywords: Pharmacovigilance, Data mining, NHIRD

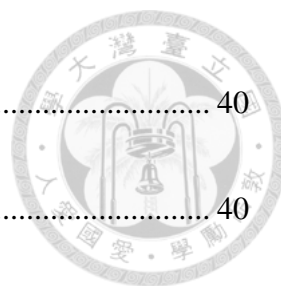
Table of Contents



誌謝	i
中文摘要	ii
ABSTRACT	iii
List of Figures.....	vii
List of Tables	x
Chapter 1 Introduction.....	1
1.1 Background.....	1
1.2 Research Motivation and Objective	4
Chapter 2 Literature Review.....	7
2.1 Spontaneous Reports Systems (SRSs)	7
2.1.1 Definition and famous examples of SRSs	7
2.1.2 Methods used on SRSs	8
2.1.3 Traits of SRSs	10
2.2 Electronic Health Records (EHR) databases	11
2.2.1 Definition and famous examples of EHR databases	11
2.2.2 Methods used on EHR databases.....	12



2.2.3 Traits of EHR databases	13
2.3 Research Gap	14
Chapter 3 Design of Our Ranking Method.....	18
3.1 Data Collection	19
3.2 Data Preparation	20
3.2.1 Data Preprocessing	21
3.2.2 Drug-Appearing Diagnosis (DAD) Generation.....	24
3.3 Learning System.....	25
3.3.1 Drug or Disease Group Mapping.....	26
3.3.2 Labeling Signals	26
3.3.3 Measure Calculation for Training Data	27
3.3.4 Summery of All Measures	34
3.3.5 Ranking Model Building	35
3.4 Detection System.....	37
3.4.1 Drug or Disease Grouping and Signal Generation.....	38
3.4.2 Measure Calculation for Candidate Signals	39
3.4.3 Rank Prediction	39



Chapter 4 Evaluation and Results.....	40
4.1 Experimental Data	40
4.2 Evaluation Design	44
4.2.1 Evaluation Criteria.....	44
4.2.2 Evaluation Procedure.....	46
4.3 Comparative Evaluation	47
4.4 Additional Evaluation	48
4.4.1 Experiment 1: Effects of Variables Selection	48
4.4.2 Experiment 2: Effects of Training Sizes.....	52
4.4.3 Experiment 3: Effects of Surveillance and Control Window Sizes.....	57
4.4.4 Experiment 4: Appropriateness of Non-Mono-Domain Training	63
Chapter 5 Conclusion and Future Work	68
References	71

List of Figures



Figure 1: Overview of learning to rank methods (Liu, 2007)	15
Figure 2: Difference of training data between pair-wise and list-wise approaches.....	16
Figure 3: Overall process of our proposed ranking method	19
Figure 4: Detailed design of the data preparation module.....	21
Figure 5: Detailed process of file linking	22
Figure 6: Drug-ATC mapping process	23
Figure 7: Example of drug-appearing diagnosis (DAD) generation	25
Figure 8: Detailed process of the learning system.....	26
Figure 9: Graphical model representation of LDA (Blei et al., 2003).....	32
Figure 10: Example of similarity between one drug and multiple diseases	34
Figure 11: Example of training data in the learning to rank method.....	36
Figure 12: Graphical view of pair-wise classification (Li, 2011).....	37
Figure 13: Detailed process of the detection system	38
Figure 14: Labeler's arrangement and their work experiences	41
Figure 15: Example of NDCG (Li, 2011).....	46
Figure 16: Comparative evaluation results (NDCG@5 to NDCG@50).....	47

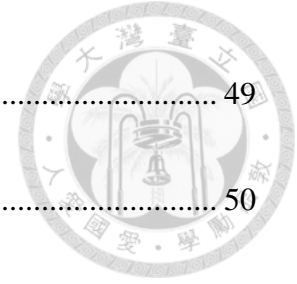


Figure 17: NDCG evaluation on type 1	49
Figure 18: NDCG evaluation on type 2.....	50
Figure 19: NDCG evaluation on type 3.....	50
Figure 20: NDCG evaluation on type 4.....	51
Figure 21: NDCG evaluation across four types of measures	51
Figure 22: NDCG evaluation for 20% training size	55
Figure 23: NDCG evaluation for 15% training size	55
Figure 24: NDCG evaluation for 10% training size	56
Figure 25: NDCG evaluation for 5% training size	56
Figure 26: NDCG evaluation across different training sizes (using Ranking SVM)	57
Figure 27: NDCG evaluation of different control window sizes.....	59
Figure 28: NDCG evaluation of different surveillance window sizes.....	59
Figure 29: NDCG evaluation of different surveillance window sizes (for hepatotoxicity)	61
Figure 30: NDCG evaluation of different surveillance window sizes (for cancer).....	61
Figure 31: NDCG evaluation of different surveillance window sizes (for cardiovascular events).....	62

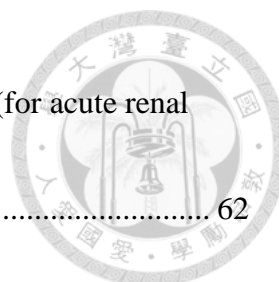


Figure 32: NDCG evaluation of different surveillance window sizes (for acute renal toxicity).....	62
Figure 33: Illustration of cross-domain training.....	64
Figure 34: NDCG evaluation for cross-domain training.....	65
Figure 35: Illustration of mixed-domain training.....	65
Figure 36: NDCG evaluation for mixed-domain training.....	67

List of Tables



Table 1: Postapproval drug withdrawals in US (Balakin, 2009; Coloma et al., 2013).....	3
Table 2: 2×2 Contingency Table for the relations between drug and event	9
Table 3: Commonly used disproportional analysis measures.....	10
Table 4: Label types and their detailed descriptions.....	27
Table 5: Contingency tables of the drug-outcome relation of {ATC-1→ICD-4}	28
Table 6: Association rule measures used in this study, where ae means adverse effect (Azevedo & Jorge, 2007).....	30
Table 7: The result table of LDA method in our study	33
Table 8: Summery of all measures used in our proposed ranking method.....	35
Table 9: Disease types and their corresponding ICD-9-CM codes	41
Table 10: Summary of our query and label collection.....	44
Table 11: Average of training and testing set in each disease query.....	46
Table 12: Comparative evaluation results (NDCG@5 to NDCG@50).....	47
Table 13: NDCG evaluation for using type 1, type 2, type 3 and type 4 measures (using Ranking SVM).....	49
Table 14: Example explaining the effect of testing size on NDCG (all ranks)	53

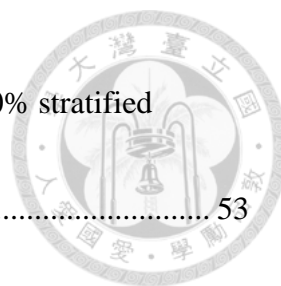


Table 15: Example explaining the effect of testing size on NDCG (50% stratified sampling of all ranks)	53
Table 16: NDCG evaluation for different training sizes.....	54
Table 17: Effects of sizes of control window and surveillance window (where c12_s12 means that control window of 12 months and surveillance window of 12 months)	58
Table 18: Number of drug-outcome pairs in each disease query under different surveillance window sizes	60
Table 19: NDCG evaluations for the cross-domain training scenario.....	64
Table 20: NDCG evaluations for mixed-domain training	66

Chapter 1 Introduction



1.1 Background

In theory, the efficacy and safety of a drug must be demonstrated during the three phases of clinical trials before approval. However, these clinical trials involve only a very limited number of participants, who may not always be representative of the population of all potential users of the drug, and cover a relatively short observation period, making it difficult to detect adverse drug reactions (ADRs) that are rare or with a long latency (Coloma, Trifirò, Patadia, & Sturkenboom, 2013; R Harpaz et al., 2012). As a result, the complete safety profile of a new drug cannot be fully established through clinical trials.

Postapproval adverse drug events (ADEs) are a global public health problem and, as Table 1 shows, many drugs were withdrawn from the market after many years of approval with harming lots of people's health. In US, there are more than 100,000 ADR-related deaths annually and cost over \$136 billion annually (Iyer, Lependu, Harpaz, Bauer-Mehren, & Shah, 2013). Similarly, it is estimated that at least 80,000 medication-related hospitalizations occur in Australia each year and more than 12,000 hospitalizations (i.e., 1.83% of all acute hospital admissions) in 2001 were related to

adverse drug reactions (ADRs) in the Netherlands (Roughead, 1999; van der Hooft, Sturkenboom, van Grootheest, Kingma, & Stricker, 2006). In the United Kingdom, ADRs account for 6.5% hospital admissions and 4% of the hospital bed capacity.

Besides, over 2% of patients admitted with an ADR died, suggesting that adverse effects may be responsible for the death of 0.15% of all patients admitted (van der Hooft et al., 2006). Therefore, to ensure the safety of public health, it is important to continue monitoring and evaluating the safety of a drug once it is on the market.

Pharmacovigilance (PhV) is defined as “the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other possible drug-related problems” (Health, 2006) with the goals of detecting novel adverse drug events earlier, reducing harms to patients, and saving social costs.

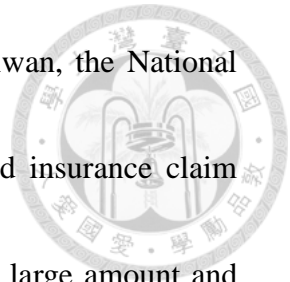
Table 1: Postapproval drug withdrawals in US (Balakin, 2009; Coloma et al., 2013)

Drug Name	Adverse Risk (Reason for Withdrawal)	Year Approved	Year Withdrawn
Cerivastatin	Phabdomyolysis	1997	2001
Rapaccuronium	Bronchospasm	1999	2001
Alosetron	Ischemic colitis	2000	2000
Cisapride	Arrhythmia	1993	1993
Phenylpropanolamine	Stroke	Pre-1962	2000
Troglitazone	Liver toxicity	1997	2000
Astemizole	Arrhythmia	1988	1999
Grepafloxacin	Arrhythmia	1997	1999
Mibefradil	Arrhythmia	1997	1998
Bromfenac	Liver toxicity	1997	1998
Terfenadine	Arrhythmia	1985	1998
Fenfluramine	Valve disease	1973	1997
Dexfenfluramine	Valve disease	1996	1997
Etretinate	Birth defects	1986	2002
Levomethadyl	Fatal arrhythmia	1993	2003
Rofecoxib	Cardiovascular events (including myocardial infarction and stroke)	1999	2004
Valdecoxib	Serious skin reactions (TENS, SJS, EM)	2001	2005
Natalizumab	Progressive multifocal leukoencephalopathy	2004	2005
Technetium fanolesomab	Cardiopulmonary failure (respiratory distress, sudden hypotension)	2004	2005
Pemoline	Liver failure	1975	2005
Pergolide	Cardiac valve damage	1998	2007
Tegaserod	Cardiovascular events (including myocardial infarction and stroke)	2002	2007
Aprotinin	Renal and cardiac complications, death	1993	2008
Efalizumab	Progressive multifocal leukoencephalopathy	2003	2009
Sibutramine	Cardiovascular events (including heart attack and stroke)	1997	2010
Gemtuzumab ozogamicin	Increased risk of death (due to liver toxicity/veno-occlusive disease)	2000	2010
Propoxyphene	Cardiac arrhythmia	1957	2010

Recently, there are two major sources for detecting ADEs, including spontaneous reports system (SRS) databases and electronic health records (EHR) databases. Some useful disproportional analysis measures have developed on SRS databases, such as RR (relative reporting), PRR (proportional reporting rate ratio), and ROR (reporting odds ratio), for ranking drug-event pairs (Iyer et al., 2013). However, the effectiveness of these analyses would be influenced by the intrinsic nature of the potential biases by reporters in volunteer and the incompleteness in spontaneous reporting.

EHR databases have provided complements for the SRS databases, because EHRs contain observational records in real world. The EHR data have potential strengths, including sufficient sample size, population basis, relative inexpensiveness, and no

possibility of recall or interviewer bias (Park et al., 2011). In Taiwan, the National Health Insurance Research Database (NHIRD) is a national-based insurance claim database, which contains claim records since March 1, 1995. This large amount and structured database contains rich information for ADR analysis, so this study focuses on developing an effective method for detecting ADRs from the NHIRD database.



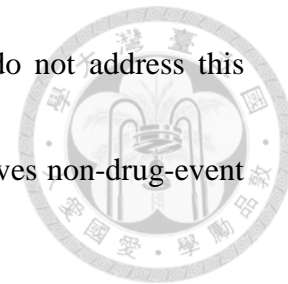
1.2 Research Motivation and Objective

Pharmacovigilance has been done for a long time, and there are several SRS systems that can be used for ADE detection. It has been successful that many adverse effects were found and some drugs with serious adverse effects were withdrawn from the market.

However, some studies show that the SRS databases cause some misleads by reporter's viewpoint and incompleteness of data (Balakin, 2009; R Harpaz et al., 2012), so we move these analyses to EHR database, NHIRD in Taiwan, for its large sample amount, quality and completeness.

Because EHRs are not collected for PhV purposes, drug-outcome pairs generated according to the time frame likely contain a vast amount of pairs that are non-drug-event pairs (e.g., drugs' indications). The existence of non-drug-event pairs undermines

detection effectiveness. However, existing EHR-based methods do not address this challenge. A novel method that attempts to identify and then removes non-drug-event pairs or decreases their importance would be desirable.



Existing EHR-based methods rely on single disproportionality analysis measures. Nevertheless, each disproportional analysis measure behaves differently to each other and may be more suitable to some situations than others. Thus, the appropriate use of multiple disproportional analysis measures may improve detection effectiveness. Furthermore, as with the SRS-based methods, all of the existing EHR-based methods are ranking-based and do not involve a supervised learning process. They simply rank drug-outcome pairs on the basis of a selected disproportional analysis measure. The use of a supervised learning method for drug safety signal detection (or surveillance) may further improve detection effectiveness.

In this study, we propose a novel EHR-based drug safety signal detection (or surveillance) method on the basis of the learning to rank approach. In addition to multiple disproportional analysis measures, the proposed method will also incorporate as candidate ranking variables 1) additional measures pertaining to the association rule research and 2) implicit relations between drugs and diseases for reducing non-drug-

event signals or decreasing their importance. We will use Taiwan's national health insurance research database covering the time-span of 2000-2009 as the data source for drug safety signal detection.



The remainder of this thesis is organized as follows. Chapter 2 reviews existing techniques related to this study, and discusses their limitations to justify our research motivation. Chapter 3 describes the data collection and our proposed method. We then present some evaluation results in Chapter 4. Chapter 5 concludes this study.

Chapter 2 Literature Review



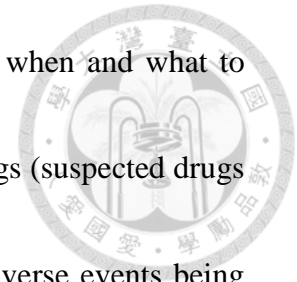
In this chapter, we review the existing databases used for pharmacovigilance and the drug safety signal detection methods pertaining to each type of the databases. In general, there are two types of the databases for pharmacovigilance: Spontaneous reporting systems (SRSs) and electronic health records (EHRs). SRSs are database resources encompassing reports of suspected post-marketed ADEs and currently represent the major data courses for pharmacovigilance (Rave Harpaz, Chase, & Friedman, 2010). In contrast to spontaneous reporting databases, EHRs databases could provide more useful information on real-world unrecognized or underappreciated drug adverse effects (AEs) (Choi, Chang, Kim, Choi, & Park, 2011). We briefly summarize these two databases, the corresponding drug safety signal detection methods, and their strengths and limitations as follows.

2.1 Spontaneous Reports Systems (SRSs)

2.1.1 Definition and famous examples of SRSs

SRSs are databases to report and save the spontaneous reports. Spontaneous reporting is dependent on potential reporters being educated and motivated to record and submit her/his observations of suspicious adverse events in voluntary. Clinicians,

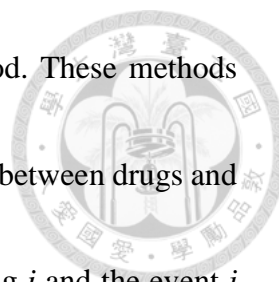
pharmacists and community members should be trained on how, when and what to report (Drug, n.d.). Each report contains patient demographics, drugs (suspected drugs and concomitant drugs) that are considered responsible for the adverse events being reported, and coded adverse events (Rave Harpaz et al., 2010).



There are two famous projects relating to SRSs for post-marketing surveillance. One is Adverse Event Reporting System (AERS) hold by The US Food and Drug Administration (FDA) and the other is World Health Organization's VigiBase. Many studies have used these two sources for detecting ADEs (Evans, Waller, & Davis, 2001; Rave Harpaz et al., 2010; Szarfman, Machado, & O'Neill, 2002).

2.1.2 Signal detection methods used on SRSs

Existing methods for detecting ADEs from SRSs generally rely on disproportional analysis measures, which detect drug-event pairs occurring at higher than expected frequencies (estimated by using information on all drugs and all events in the database) (Almenoff et al., 2007; Coloma et al., 2013; R Harpaz et al., 2012; Lependu, Iyer, Fairon, & Shah, 2012). Common disproportional analysis measures (i.e., measures of association) include the relative reporting ratio (RRR), proportional reporting ratio (PRR), reporting odds ratio (ROR), and information component (IC) calculated by the



Bayesian confidence propagation neural network (BCPNN) method. These methods use the 2×2 contingency table (see Table 2) to describe the relations between drugs and events. It summarizes the number of reports that have the focal drug i and the event j of interest as “a”, the focal drug and other events as “b”, other drugs and the target event j as “c”, and other drugs and other events as “d”. A similar table is constructed for every possible drug-event combination (Hauben & Bate, 2009).

Table 2: 2×2 Contingency Table for the relations between drug and event

	Event j	Other Events	Total
Drug i	a	b	a+b
Other Drugs	c	d	c+d
Total	a+c	b+d	a+b+c+d

By using the information in the contingency table, the disproportional analysis measures mentioned above compare the observed counts of drug–event relations with the expected counts based on the relative frequency of events occurring for the drug alone and the event alone (Choi et al., 2011). Detailed formulas are listed in Table 3. The more the number of the observed reports exceeds the number of expected reports by chance, the more interesting, possibly and worthy for further investigation (Hauben & Bate, 2009).

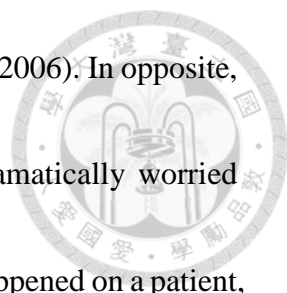
Table 3: Commonly used disproportional analysis measures

Disproportional analysis measure	Formula	Probabilistic interpretation
RRR	$\frac{a(a + b + c + d)}{(a + c)(a + b)}$	$\frac{Pr(ae drug)}{Pr(ae)}$
PRR	$\frac{a(c + d)}{c(a + b)}$	$\frac{Pr(ae drug)}{Pr(ae \sim drug)}$
ROR	$\frac{ad}{cb}$	$\frac{Pr(ae drug)Pr(\sim ae \sim drug)}{Pr(\sim ae drug)Pr(ae \sim drug)}$
IC	$\log_2 \frac{a(a + b + c + d)}{(a + c)(a + d)}$	$\frac{\log_2 Pr(ae drug)}{Pr(ae)}$

2.1.3 Traits of SRSs

The voluntary nature of spontaneous reports makes SRS-based drug safety surveillance system reactive or even passive. In addition, there are some inherent limitations existing in SRS databases such as underreporting, overreporting, duplicate reporting, misattribution of causality in drug–event combinations, missing or incomplete data and not enough data for denominator (Balakin, 2009; R Harpaz et al., 2012).

Underreporting is the major problem of SRS data for only about 10% of serious adverse events are reported by Hazell & Shakir’s study (2006). The reasons for not reporting include a lack of time, different care priorities, uncertainty about the drug causing an ADR, difficulty in accessing reporting forms, lack of awareness of the requirements of the purpose of SRSs, less likely to report well-known and trivial ADRs,



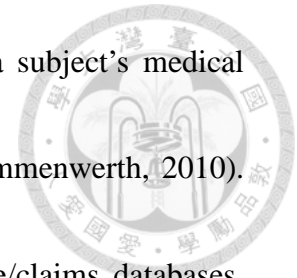
and physicians' attitudes towards reporting ADRs (Hazell & Shakir, 2006). In opposite, some serious ADRs are overreporting by media attention and dramatically worried about the ADRs, causing many spurious data. Besides, if one case happened on a patient, this case may be alerted to the SRS by the patient and his attending doctor and nurses. When all of them did reporting, the duplicate reporting happened. Also, because of the SRSs are reported by human beings, they must have some personal view in the reports that may cause the misattribution of causality in drug–event combinations. The data in SRS database may be incomplete or missing some attributes because the reporters could be the patients which do not have enough knowledge. Last, the SRSs are lack of enough denominator cases needed in disproportional analysis measures, so the significant relations picked up may not really significant in real world (Choi et al., 2011). To address the limitations of SRS-based drug safety surveillance systems, several research initiatives have been carried out to explore the use of EHRs for developing active surveillance systems.

2.2 Electronic Health Records (EHR) databases

2.2.1 Definition and famous examples of EHR databases

EHR, especially nation-wide health insurance claims databases, are population-

based. It includes data that is not only particularly relevant to a subject's medical treatment but also to a subject's health in general (Hoerbst & Ammenwerth, 2010).



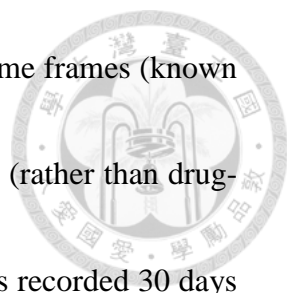
EHR includes either medical records databases or administrative/claims databases.

Medical records databases, which include records maintained for the management of patients' clinical care; whereas administrative/claims databases, which include transactions primarily to achieve administrative purposes, such as claims for reimbursement from insurance companies (Strom, 2012).

Initiatives like the Observational Medical Outcomes Partnership (OMOP) in the US and the Exploring and Understanding Adverse Drug Reactions (EU-ADR) project in Europe focus on building EHR-based drug safety surveillance systems. These projects mainly utilize electronic medical records or administrative/claims databases for identifying drug adverse reactions (Iyer et al., 2013). In Taiwan, the national insurance claims database covers almost all of Taiwanese. Therefore, it is believed that the large amount and structured data would provide a lot of information for post-marketing drug safety surveillance.

2.2.2 Signal detection methods used on EHR databases

In EHR databases, there are no direct drug and ADE connections. Existing EHR-

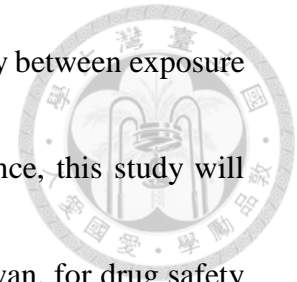


based methods generally use the temporal information to identify time frames (known as surveillance windows or hazard periods) in which drug-outcome (rather than drug-event) pairs are identified and analyzed, e.g., outcomes or diagnoses recorded 30 days after drug exposure (R Harpaz et al., 2012). Then, a prevalent approach for EHR-based drug safety surveillance adopts and extends the disproportional analysis measure commonly employed by existing SRS-based methods which use a specific disproportional analysis measure for signal detection (i.e., ranking drug-outcome pairs using a selected disproportional analysis measure). In this way, these measures may enable the identification of real signals that were missed from the SRS databases due to incorrect records and underreporting (Reps, Feyereisl, & Garibaldi, 2011).

2.2.3 Traits of EHR databases

In EHRs, the data are real cases and real time records by professional physicians, so it avoids the problems associated with SRSs. EHR-based drug safety surveillance systems rely on data collected from routine clinical care rather than voluntary. Thus, their signal detection endeavors can proceed actively rather than passively or reactively. In addition, the large quantity of the patients' records in EHRs provide more precise denominator fitted in real world; the longitudinal nature of routinely-collected EHR

may allow the identification of adverse events that have a long delay between exposure and clinical manifestations (Coloma et al., 2013). As a consequence, this study will employ the NHIRD, the national insurance claims database in Taiwan, for drug safety signal detection.

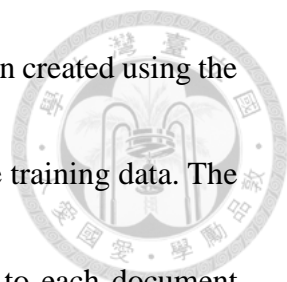


2.3 Research Gap

The surveillance ability of SRS-based measures is restricted due to the nature of unveiling reports spontaneously. Thus, this leads to the use of electronic health records, which is routinely-collected by healthcare institutes, for pharmacovigilance.

Existing EHR-based surveillance systems are also based on a specific disproportional analysis measure for signal detection, and rank drug-outcome pairs by this specific measure. However, as we mentioned previously, different measures may be suited to different situations. Thus, we tend to combine the traits of multiple measures by the learning to rank approach, and we believe that this could improve the effectiveness of signal detection.

As Figure 2 illustrates, learning to rank methods often separate into two processes, which together deal with the ranking problem. The first process is learning. In the learning system, a number of queries (q_n) are provided, where each query is associated



with a perfect ranking list ($y^{(n)}$) of documents; a ranking model is then created using the training data, such that the model can predict the ranking lists in the training data. The other part is the ranking system, which assigns a relevance score to each document pertaining to a given query and ranks the documents in the descending order (Z. Cao, Qin, Liu, Tsai, & Li, 2007).

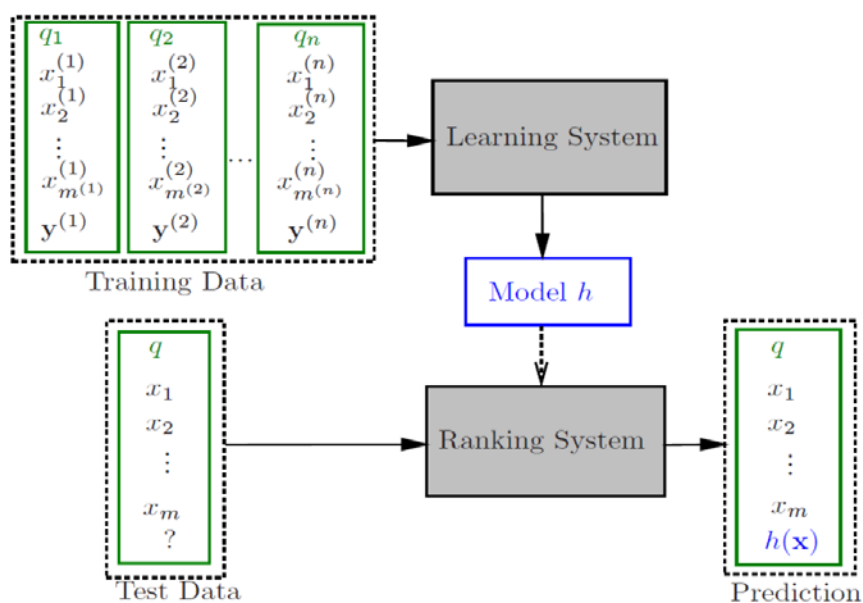


Figure 1: Overview of learning to rank methods (Liu, 2007)

Besides, there are two major approaches for learning to rank: pair-wise and list-wise. As Figure 2 shows, in pair-wise methods, the query-based document groups will be broken into lots of document pairs with higher or lower ranks, so one query's relating document group would become a lot of document ranking pairs as training data. In opposite, the list-wise methods use the whole query-based document list groups as training data, so the group structure of ranking is maintained.

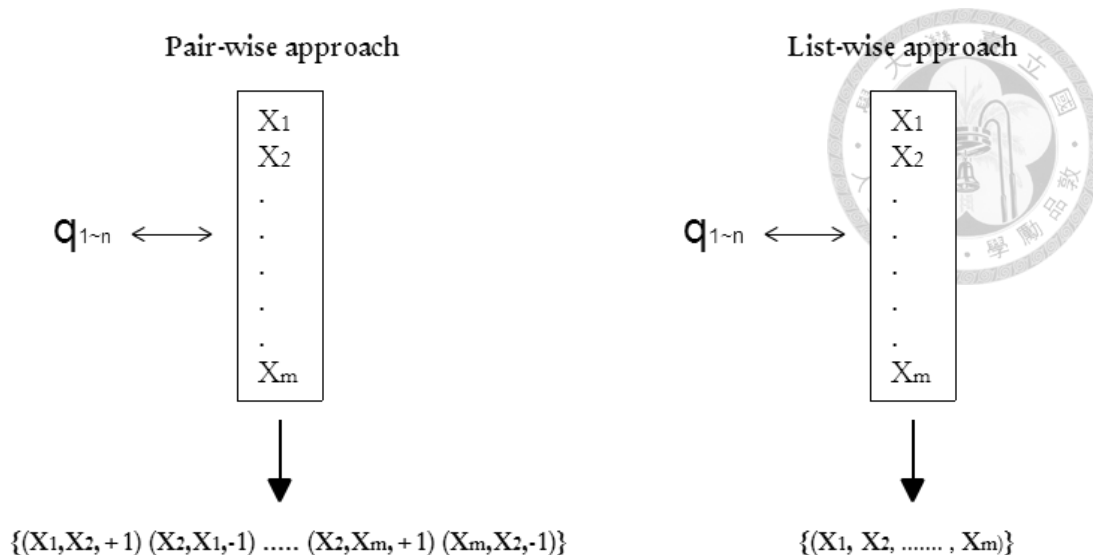


Figure 2: Difference of training data between pair-wise and list-wise approaches

In our study, there are two kinds of queries that can be used in training and testing. One is drug-anchored drug-outcome pairs and the other is disease-anchored drug-outcome pairs. The definition of drug-anchored drug-outcome pairs is, given one drug as the detection target (query), the relations between this drug and diseases possibly caused by this drug form as drug-anchored drug-outcome pairs (documents). The disease-anchored drug-outcome pairs behave similarly. Given one or a group of diseases as the detection target (query), all the drugs possibly causing this disease or disease group constitute disease-anchored drug-outcome pairs (documents). To prepare a training data set for either drug-anchored or disease-anchored drug-outcome pairs, the ranking of drug-outcome pairs within a query has to be labeled by professional pharmacists. This labeling process costs lots of effort and time. Thus, the number of

queries that can be collected will be very limited, making the use of the list-wise approach infeasible. Consequently, we will adopt the pair-wise learning to rank approach in this study.



Chapter 3 Design of Our Proposed Ranking Method



As mentioned previously, we attempt to detect candidate adverse drug effects from the NHIRD and rank these candidate adverse drug effects by our proposed ranking method. As Figure 3 illustrates, our proposed ranking method consists of three main modules, including data preparation, learning system, and detection system. In the first module (i.e., data preparation), we extract useful data from the NHIRD database and perform the data preprocessing for every patient's records according to a prespecified sizes of control window and surveillance window. Second, in the learning system, we construct a signal ranking model from one or some lists of labeled signals (i.e., training data set) corresponding to a specific detection target (drug-anchored or disease-anchored). Finally, given a detection target, the detection system generates all candidate drug-outcome pairs and ranks these signals on the basis of the single ranking model built by the learning system. In the following, we first describe the data collection (i.e., NHIRD database) used in this study. Subsequently, we depict the detailed design of each module in our proposed ranking method.

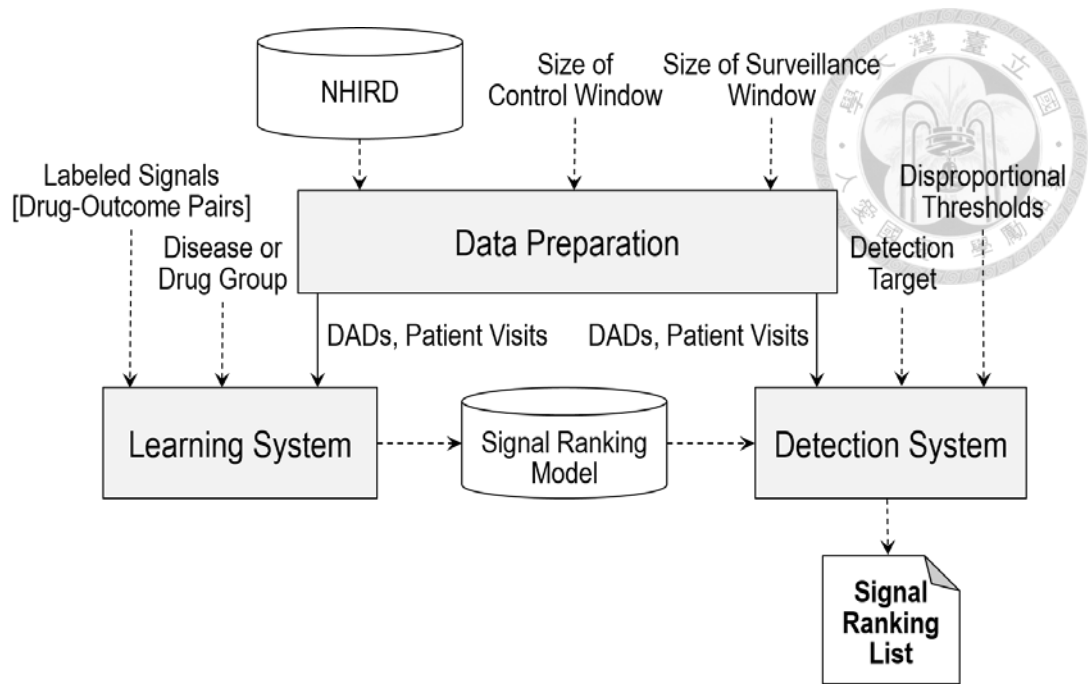


Figure 3: Overall process of our proposed ranking method

3.1 Data Collection

The NHIRD, a large-scale computerized database, collected by the Bureau of NHI and maintained by the National Health Research Institutes (NHRI), is provided to scientists in Taiwan for research purposes (Lin et al., 2014). In NHIRD, there are registration files correspond to medical records and original claim data. NHRI randomly selects one million patients and their whole relating records in registration files and original claims data. In this research, we only use the claims data including the files of DD, CD, GD, GO, OO, and DO. DD, CD and GD files contain the patient's visit data and diagnoses; GO, OO, and DO contain the prescription, including drugs.

The following is the descriptions of these six files of claim data:

1. DD: Inpatient expenditures by admissions.
2. CD: Ambulatory care expenditures by visits.
3. GD: Expenditures for prescriptions dispensed at contracted pharmacies.
4. DO: Details of inpatient orders.
5. OO: Details of ambulatory care orders.
6. GO: Details of prescriptions dispensed at contracted pharmacies.



3.2 Data Preparation

In the data preparation module, as demonstrated in Figure 4, there are two steps to transform the original data into the patient visits and drug-appearing diagnoses (DADs) that will be used for calculating the measurements.

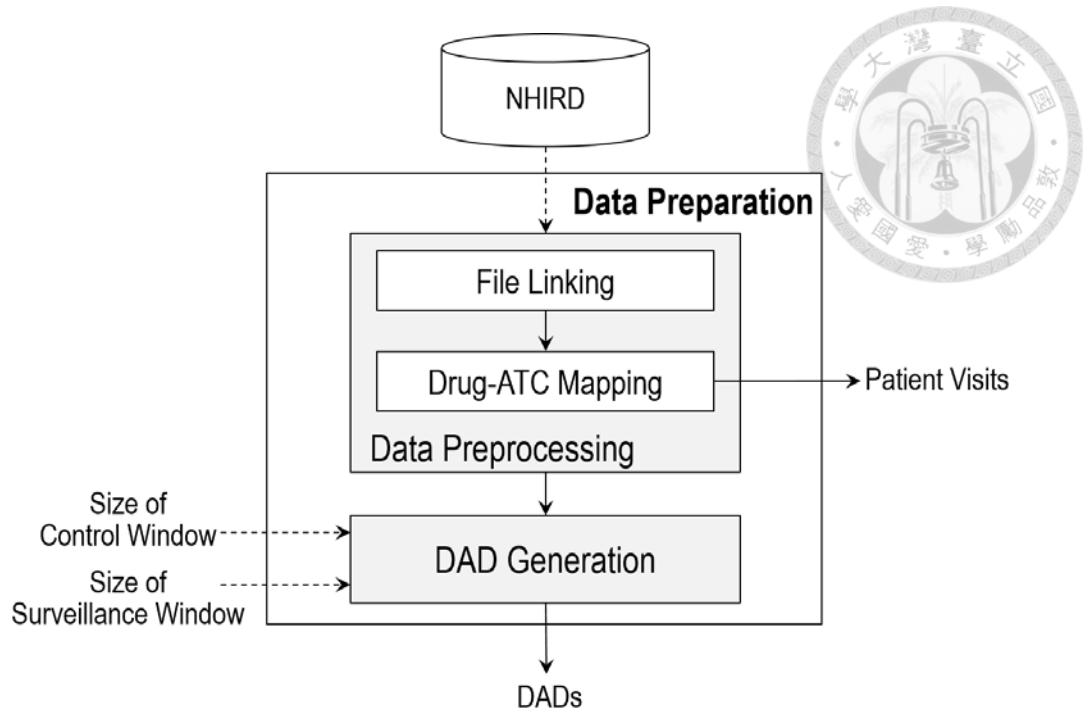


Figure 4: Detailed design of the data preparation module

3.2.1 Data Preprocessing

As we mentioned above, patient’s diagnosis and drugs are saved separately in two different types of files. As Figure 5 shows, we combine GD and GO, CD and OO, and DD and DO by the same foreign keys and store the information about each patient visit such as patient id, visit time, diagnoses, drugs prescribed and so on into one file which we called “Patient visits.” Thus, we can get every patient visit data completely.

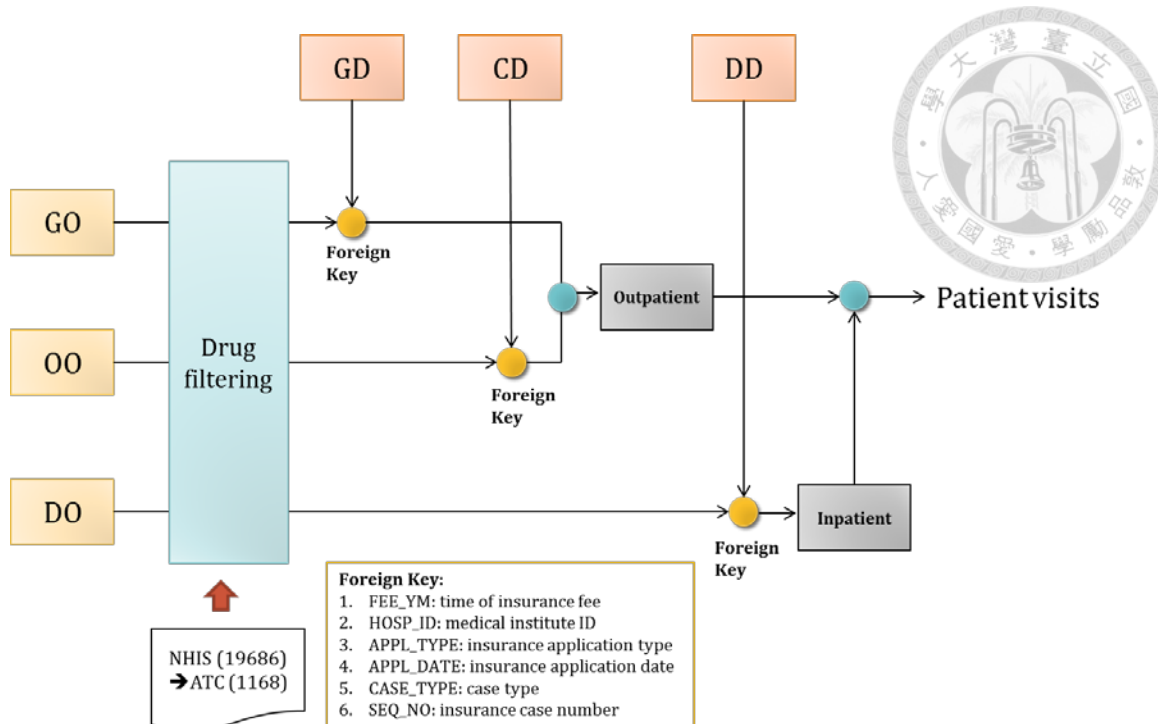


Figure 5: Detailed process of file linking

Regarding drug filtering illustrated in Figure 5, drugs recoded in NHIRD are codes called National Health Insurance Scheme (NHIS) Medicines List, which classify the drugs by the generic name, the amount of dosage and the ways to take medicines for easy pricing. However, one generic name drug may contain not only one component, and different dosage and taking medicine way would be named as different codes. This makes us difficult to learn the associations between drugs and diseases. Thus, we recruited graduate students of School of Pharmacy at National Taiwan University and pharmacists in National Taiwan University Hospital to establish the mapping between the drugs in the NHIS Medicines List and their Anatomical Therapeutic Chemical (ATC) codes, which is used worldwide for classifying medical substances and serves as a tool

for drug utilization research. This classification system divides drugs into different groups according to the organ or system on which they act and/or their therapeutic and chemical characteristics (Chen, Zeng, Cai, Feng, & Chou, 2012). The WHO recommends the ATC system for international comparisons, and it is also used for reporting of adverse drug reactions¹. Figure 6 illustrates the detail of drug-ATC mapping. Accordingly, the drug-ATC mapping contains 19,686 NHIS drugs with 1,168 ATC codes. Hence, the NHIS codes in the patient-visit file are substituted by their ATC codes. Regarding diseases, the NHIRD uses International Classification of Diseases (ICD-9-CM), the official system of assigning codes to diagnoses and procedures associated with hospital utilization ((US), 1980), for recording diseases, so there is no need for transforming.

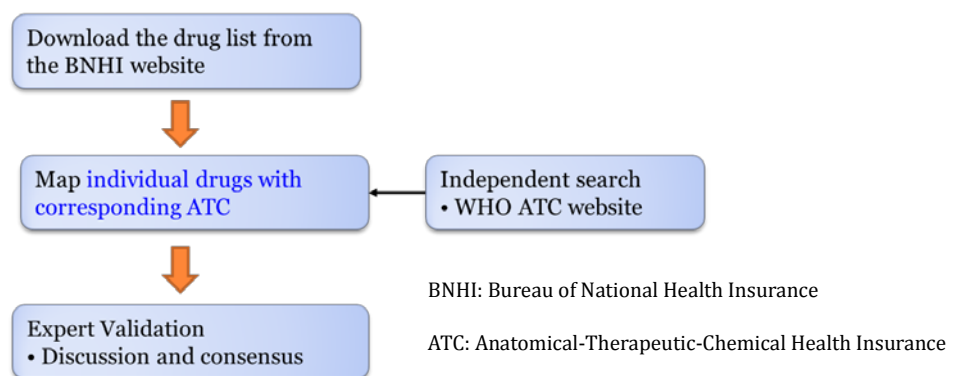


Figure 6: Drug-ATC mapping process

¹ This information is retrieved from: <http://sydney.edu.au/medicine/fmrc/atc/index.php>.

3.2.2 Drug-Appearing Diagnosis (DAD) Generation



From patient-visit data, we place every visit of one patient on a timeline by time order, and then extract the DAD relations by the following steps. First, as Figure 7 illustrates, for patient visit V_i , we extract the current diagnoses (i.e., ICD-1 and ICD-2), the post diagnoses in the surveillance window T_s (i.e., ICD-2, ICD-3, and ICD-4), and the preexisting conditions in the control window T_c (i.e., ICD-1 and ICD-3). Please note that we exclude ICD codes related to physical examinations. Second, to avoid those events before the drug event (i.e., ATC-1, ATC-2) such as chronic diseases, we obtain the appearing diagnoses (i.e., ICD-4) by removing the diagnoses in the preexisting conditions and the current diagnoses from the post diagnoses. Then, the drug-appearing diagnosis (DAD) of V_i (i.e., {ATC-1, ATC-2; ICD-4}) contains the drugs prescribed in the V_i (i.e., ATC-1 and ATC-2) and the appearing diagnoses (i.e., ICD-4) obtained from the previous steps. Finally, we construct drug-appearing diagnoses (DAD) for each patient visit in the duration between year 2000 to 2009 across all patients, and then calculate the following measures on these visit-based DADs.

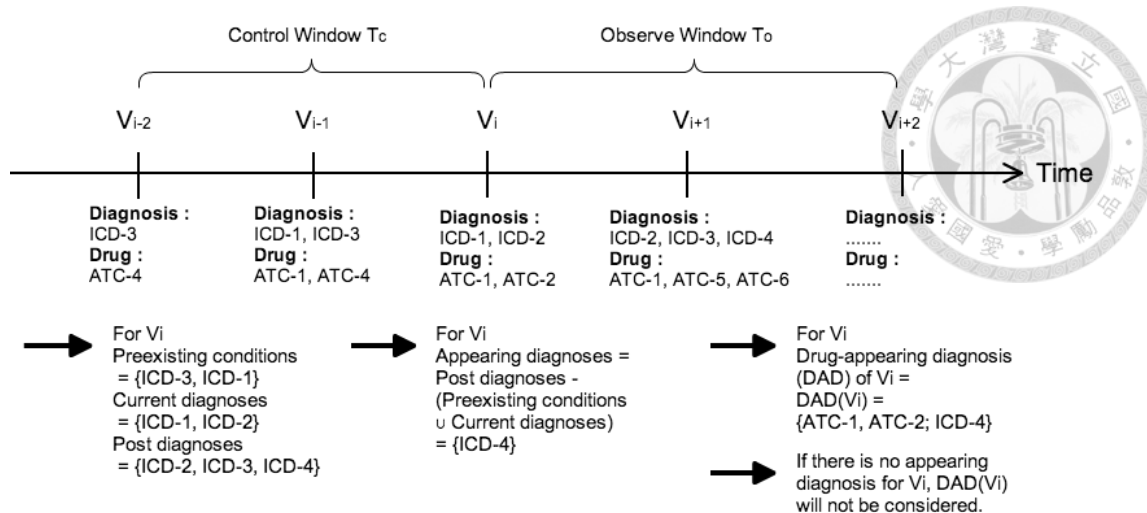


Figure 7: Example of drug-appearing diagnosis (DAD) generation

3.3 Learning System

In the learning system, we use Ranking SVM, a pair-wise learning to rank method, to build our signal ranking model. Before building model, the ATC or ICD code in DADs are mapped to drug or disease group, and then we combine DADs and patient visits with labeled signals (drug-outcome pairs) to calculate measures for the training data. Notice that, the labeled signals could be a small proportion of one/a group of drug(s) with related diseases (drug-anchored) or one/a group of disease(s) with related drugs (disease-anchored). The process of the learning system is presented in Figure 8.

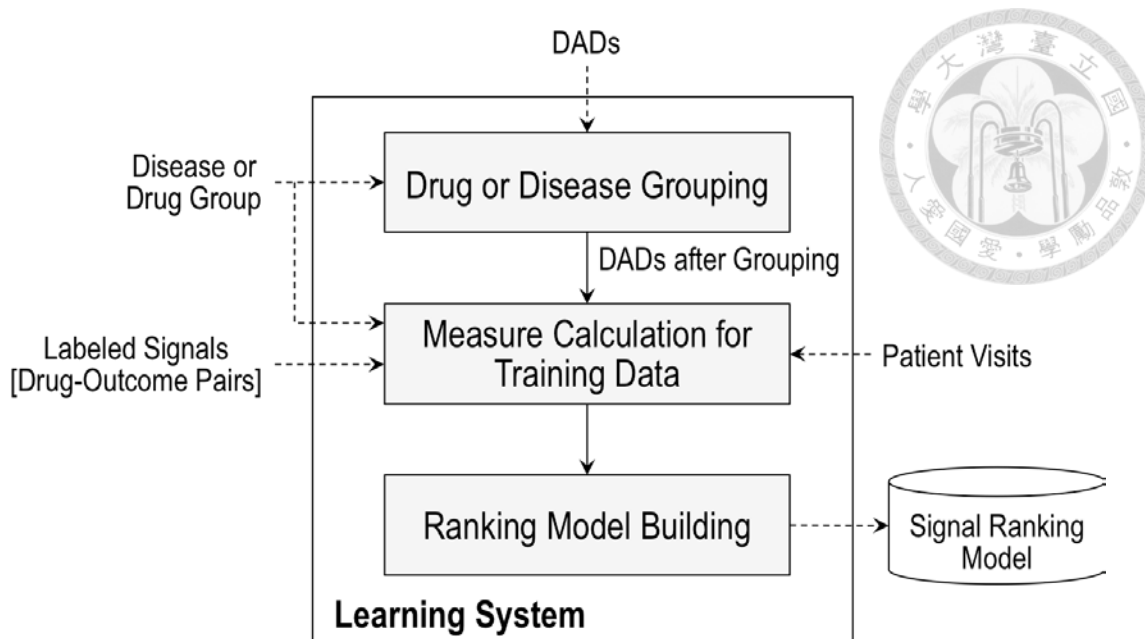


Figure 8: Detailed process of the learning system

3.3.1 Drug or Disease Group Mapping

In this step, we map ATC or ICD codes in the DADs to the predefined drug or disease group of interest. If we use a drug group as our detection target, then the training and testing data are drug-anchored DADs. The disease group behaves in the same way, and we call such DADs as disease-anchored DADs. For example, the acute renal toxicity is a disease group consisting of ICD codes of 584 and 586, so we map the ICD codes of 584 and 586 in the DADs to acute renal toxicity.

3.3.2 Labeling Signals

Experts follow the rules in Table 4 to label the drug-disease pairs relative to disease or drug group as the training data. Each label type corresponds to a level of likelihood.

Our study ranks these pairs by the likelihood from high (4) to none (1), and these labeled data will be used for training and testing purposes.



Table 4: Label types and their detailed descriptions

Label Type	Description	Likelihood
Known ADEs	This event has information of package inserts from manufacturers (Product information) or epidemiological study, ex. population-based cohort study, case-control study, hospital-based study.	High
Suspected ADEs	This event has some series case reports or animal-case study, in vivo and in vitro experimental study.	Medium
Unknown associations	This event doesn't appear in the known ADEs, possible ADEs or Indication associations, but it may or may not be detected as an ADE in the future.	Low
Indication associations	An "indication" for a drug refers to the use of that drug for treating a particular disease, so this event is not a drug-ADE pair.	None

3.3.3 Measure Calculation for Training Data

There are three types of measures in our study, which are traditional disproportional analysis measures, other association rule measures and drug-disease association measure. Traditional disproportional analysis measures and other association rule measures are based on the contingency table shown in Table 2. Here, we show an example of constructing the contingency table. If there are three DADs which are {ATC-1, ATC-2; ICD-4}, {ATC-1, ATC-3; ICD-4} and {ATC-4; ICD-3},

and the contingency tables of the drug-outcome relations of {ATC-1→ICD-4} are shown in Table 5 by calculating the number of DADs. Differently, drug-disease association measures use patient-visit data illustrated in Chapter 3.2.1 to calculate this measure.

Table 5: Contingency tables of the drug-outcome relation of {ATC-1→ICD-4}

	ICD-4	No ICD-4
ATC-1	2	0
No ATC-1	0	1

1. Traditional disproportional analysis measures

Existing EHR-based drug safety surveillance methods adopt and extend the disproportional analysis measures commonly employed by SRS-based methods, such as RRR, PRR, ROR and BCPNN. RRR, PRR and ROR are illustrated elaborately in Chapter 2, so we only explain how we calculate BCPNN here.

According to Lindquist & Olsson (1998), the BCPNN estimates the information component (IC), which is based on the definition of RRR. The information component (IC) is the strength of the association between two variables and is the logarithmic form of the symmetrical factor relating to the prior and posterior probability.

$$IC = \log_2 \frac{P(\text{drug } i, \text{ADR } j)}{P(\text{drug } i) P(\text{ADR } j)}$$

Because we do not know the “real” probabilities of $P(\text{drug } i)$, $P(\text{ADR } j)$, and $P(\text{drug } i, \text{ADR } j)$, we assert a beta distribution for each probability. From these distributions we calculate the “expectation values” of the beta distribution of each variable as BCPNN. Therefore, the BCPNN is calculated by:

$$E(IC_{ij}) = \log_2 \frac{E(P_{ij})}{E(P_i) E(P_j)} = \log_2 \frac{\left(\frac{N_{ij} + \gamma_{11}}{N + \gamma}\right)}{\left(\frac{N_i + \alpha_1}{N + \alpha}\right) \cdot \left(\frac{N_j + \beta_1}{N + \beta}\right)}$$

where i, j corresponding to drug i and ADR j , N is the total number, α_1 and α_0 are the factors in the beta distribution of $P(\text{drug } i)$ and $P(\text{ADR } j)$, and γ_{11} and γ are the corresponding factors for the joint probability $P(\text{drug } i, \text{ADR } j)$. The priori probability of $P(\text{drug } i)$ and $P(\text{ADR } j)$ are assumed equal, because any probability is the same as any other without further information; in a beta distribution this corresponds to the constants α_1 and α_0 (where $\alpha = \alpha_1 + \alpha_0$) and $\alpha_1 = \alpha_0 = 1$. γ_1 and γ define the joint beta distribution $P(\text{drug } i, \text{ADR } j)$. We set $\gamma_{11} = 1$ and define γ as:

$$\gamma = \frac{\gamma_{11}}{P(\text{drug } i) P(\text{ADR } j)}$$

2. Other association rule measures

In this section, we describe the association rule measures used in our study in the following table.

Table 6: Association rule measures used in this study, where ae means adverse effect (Azevedo & Jorge, 2007)



Measure	Formula	Probabilistic interpretation	Description
Confidence	$\frac{a}{a+b}$	$\frac{\Pr(\text{drug} \cap \text{ae})}{\Pr(\text{drug})}$	Confidence is an estimate of $P(\text{ae} \text{drug})$ and ranges from 0 to 1.
Conviction	$\frac{(a+b)(b+d)}{b(a+b+c+d)}$	$\frac{1 - \Pr(\text{ae})}{1 - \text{confidence}}$	Conviction is sensitive to rule direction ($\text{conv}(\text{drug} \rightarrow \text{ae}) \neq \text{conv}(\text{drug} \leftarrow \text{ae})$). It could capture the notion of implication rules and ranges from 0.5 to ∞ .
Leverage	$\frac{ad - bc}{(a+b+c+d)^2}$	$\Pr(\text{drug} \cap \text{ae}) - \Pr(\text{drug}) \Pr(\text{ae})$	Leverage is to measure how much more counting is obtained from the co-occurrence of the drug and ae from expected. It ranges from -0.25 to 0.25.
χ^2	$N \times \sum_{\substack{X \in (a+b, c+d) \\ Y \in (a+c, b+d)}} \frac{((X \cap Y) - \frac{X \cdot Y}{N})^2}{X \cdot Y}$	$N \times \sum_{\substack{X \in (\text{drug}, \neg \text{drug}) \\ Y \in (\text{ae}, \neg \text{ae})}} \frac{((X \cap Y) - \frac{X \cdot Y}{N})^2}{X \cdot Y}$	χ^2 is the definite way for measuring the statistical independence between drug and ae, and the value doesn't related to the correlation strength. It ranges from 0 to ∞ .
Jaccard	$\frac{a}{a+b+c}$	$\frac{\Pr(\text{drug} \cap \text{ae})}{\Pr(\text{drug}) + \Pr(\text{ae}) - \Pr(\text{drug} \cap \text{ae})}$	Jaccard coefficient assesses the distance between drug and ae. Higher value indicates that the overlap between drug and ae is more. It ranges from 0 to 1.
Cosine	$\frac{a}{\sqrt{(a+b)(a+c)}}$	$\frac{\Pr(\text{drug} \cap \text{ae})}{\sqrt{\Pr(\text{drug}) \Pr(\text{ae})}}$	Cosine is another way to measure distance drug and ae on vector space. It ranges from

			0 to 1.
Φ-coefficient	$\frac{ad - bc}{\sqrt{(a + b)(a + c)(c + d)(b + d)}}$	$\frac{leve}{\sqrt{\Pr(\text{drug}) \Pr(\text{ae}) \Pr(\neg \text{drug}) \Pr(\neg \text{ae})}}$	Φ -coefficient measures the association between drug and ae by Pearson correlation coefficient. It ranges from -1 to 1.

3. Drug-disease association measures

In drug-disease association extraction, we want to find the implicit relations between drugs and diseases. Latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003) is a generative probabilistic model of a corpus. This method could figure out the possible relations of drugs and diseases by the probability of distributions of each drug and disease over topics. The basic idea of LDA is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. Relating to our study, the drugs and diseases that often occur together (in this case, the diseases are often the indications of the drugs) would have large probability to be distributed in the same topics. LDA assumes the following generative process for each document w in a corpus D :

1. Choose $N \sim \text{Poisson}(\xi)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the N words w_n :



- (a) Choose a topic $z_n \sim \text{Multinomial}(\theta)$.
- (b) Choose a word w_n from $p(w_n | z_n, \beta)$, a multinomial probability conditioned on the topic z_n .

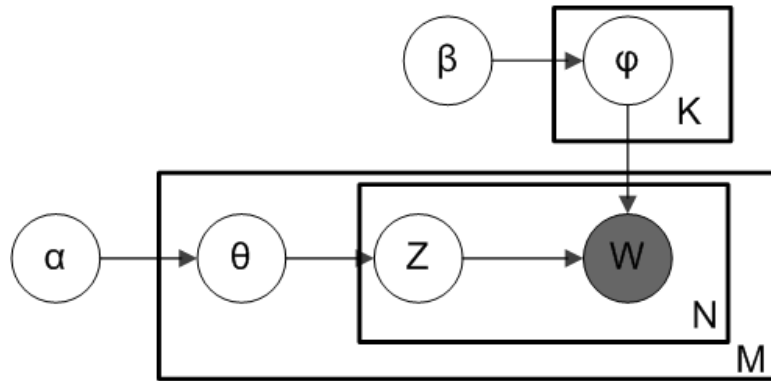
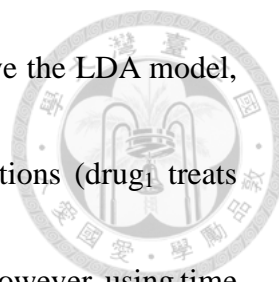


Figure 9: Graphical model representation of LDA (Blei et al., 2003)

The probability of a sequence of words and topics are calculated by following formula, and a brief structure of this formula is showed in Figure 9. The word probabilities are parameterized by a $k \times V$ matrix β where $\beta_{ij} = p(w_j = 1 | z_i = 1)$, and k is the number of topics we choose and V is the number of words in the whole corpus.

$$p(w, z) = \int p(\theta) \left(\prod_{n=1}^N p(z_n | \theta) p(w_n | z_n) \right) d\theta$$

In this study, we believe that some possible implicit relations exist between drugs and diseases. For example, one patient go to hospital three times in sequence in a time window, and the diagnosis and relating drugs are visit₁: diagnosis₁, drug₁, visit₂: diagnosis₂, drug₂ and visit₃: diagnosis₃, drug₃. We use the patient-visit file which



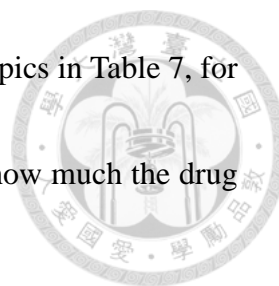
contains every visit (time window = 0 day) of each patient to derive the LDA model, and then we may extract out some relations such as drug indications (drug₁ treats diagnosis₁) if there are many visits contained drug₁ and diagnosis₁. However, using time window = 0 day may be too specific to capture the drug-disease implicit relations, so we also extend the time window to 30 days. We hope that this method could have the ability to capture additional relations such as drug progression that disease₁ progressed to disease₂ and disease₃ within 30 days. In this case, if disease₁ is an indication of drug₁, then both the disease₂ and disease₃ are unlikely to be the adverse effects of drug₁.

Specifically, we use Gibbs sampling LDA package² to capture the implicit relations between drugs and diseases. The number of topics we choose are 50 and 75, and the hyperparameters α and β were set to 0.5 and 0.1.

Table 7: The result table of LDA method in our study

Words	Topic 1	Topic 2	Topic k
ATC-1	P(ATC-1,Topic 1)	P(ATC-1,Topic 2)	P(ATC-1,Topic k)
ATC-2	P(ATC-2,Topic 1)	P(ATC-2,Topic 2)	P(ATC-2,Topic k)
.....
ICD-1	P(ICD-1,Topic 1)	P(ICD-1,Topic 2)	P(ICD-1,Topic k)
ICD-2	P(ICD-2,Topic 1)	P(ICD-2,Topic 2)	P(ICD-2,Topic k)
.....

² GibbsLDA++: <http://gibbslda.sourceforge.net>



To utilize the probability of the disease and the drug over all topics in Table 7, for every drug-outcome pair, we use the cosine similarity to calculate how much the drug and the disease is related. The cosine similarity's formula is:

$$\text{cosine similarity} = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

where A is the drug's probability vector over n topics and B is the disease's probability vector over n topics. Besides, for drug-anchored or disease-anchored drug-outcome pairs, there would be many cosine similarity values related to this drug or disease group. Thus, as Figure 10 shows, we choose the maximum of these cosine similarities as the overall similarity between one drug and many diseases.

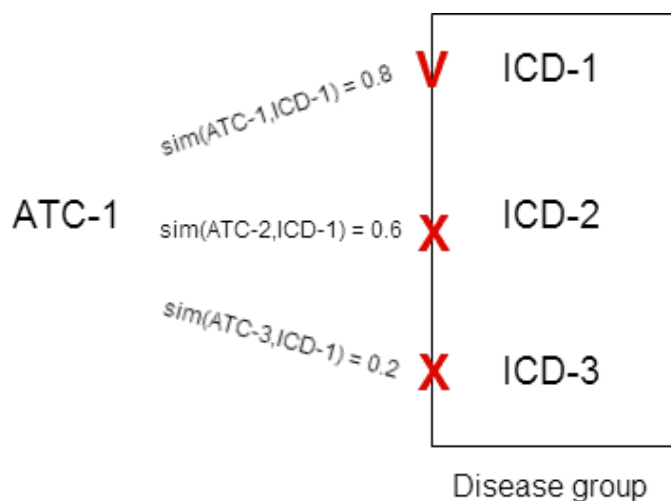


Figure 10: Example of similarity between one drug and multiple diseases

3.3.4 Summary of All Measures

The following table is the summary of our measures used in our proposed

ranking method.

Table 8: Summary of all measures used in our proposed ranking method

Types of Measures	Measures
Traditional disproportional analysis measures	RRR
	PRR
	ROR
	BCPNN
Other association rule measures	Confidence
	Conviction
	Leverage
	χ^2
	Jaccard
	Cosine
	Φ -coefficient
Disease-drug association measures	Disease-drug association: window=0, topic=50
	Disease-drug association: window=0, topic=75
	Disease-drug association: window=30, topic=50
	Disease-drug association: window=30, topic=75

3.3.5 Ranking Model Building

As mentioned in Figure 8, we combine the signals (drug-outcome pairs) labeled by experts with all the measures illustrated above as the training data and then use a pair-wise learning to rank method to construct a signal ranking model. In our study, the queries mapped to the disease or drug group, and the documents mapped to disease-anchored or drug-anchored drug-outcome pairs (see Figure 11).

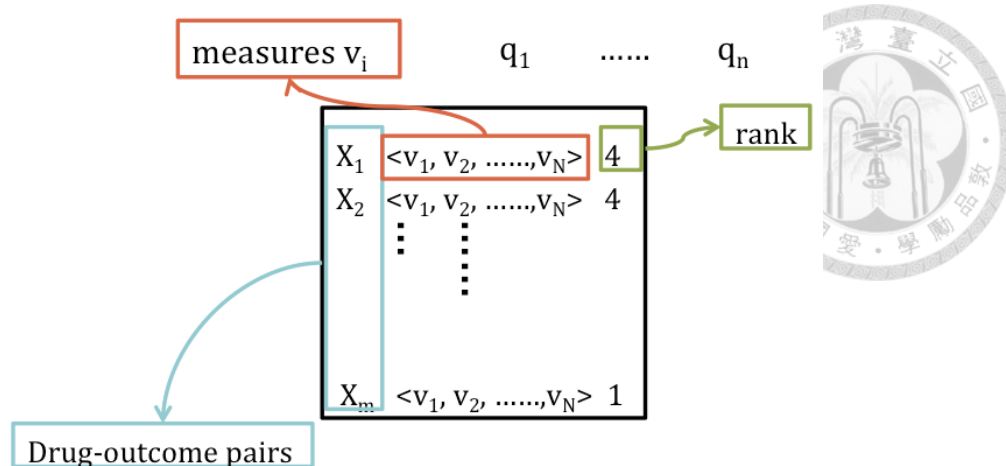


Figure 11: Example of training data in the learning to rank method

Ranking SVM³ proposed by Herbrich et al (1999) is a popular pair-wise learning to rank method (Y. Cao et al., 2006). This method forms a ranking model by minimizing a regularized margin-based pair-wise loss. The queries shown in Figure 2 are the input data and the objective function is:

$$\min_{\omega, \xi} \frac{1}{2} \|w\|^2 + C \xi_i$$

$$\text{subject to: } Z_i \langle w, x_i^{(1)} - x_i^{(2)} \rangle \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

$$i = 1, \dots, N$$

where $x_i^{(1)}$ and $x_i^{(2)}$ denote the first and second feature vectors in a pair of feature vectors, and N is the number of training instances. The constant $C > 0$ is the trade-off parameter between the margin size and the amount of errors. The slack variables ξ_i

³ SVM light: http://www.cs.cornell.edu/people/tj/svm_light/index.html

measure the degree of misclassification. The objective function is to optimize a solution vector w for ranking model (Yu & Kim, 2012). Figure 12 shows a brief graphical view of this pair-wise method. The pair of $X_1 - X_3 = +1$ means that X_1 ranks higher than X_3 , and the opposite situation of $X_3 - X_1 = -1$ means that X_3 ranks lower than X_1 . Ranking SVM tries to find a solution vector w that can classify these pairs correctly. In our study, we use the RankSVM in SVMLight (Chapelle & Keerthi, 2009) to build a signal ranking model.

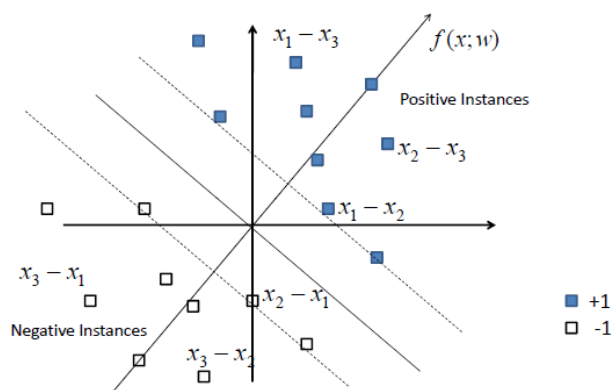


Figure 12: Graphical view of pair-wise classification (Li, 2011)

3.4 Detection System

In the detection system, we prepare the drug-anchored or disease-anchored drug-outcome pairs as testing data by the signal generation process and subsequently perform the measure calculation for candidate signals. We then put them into the signal ranking

model for rank prediction. Figure 13 illustrates the detailed process of the detection system.

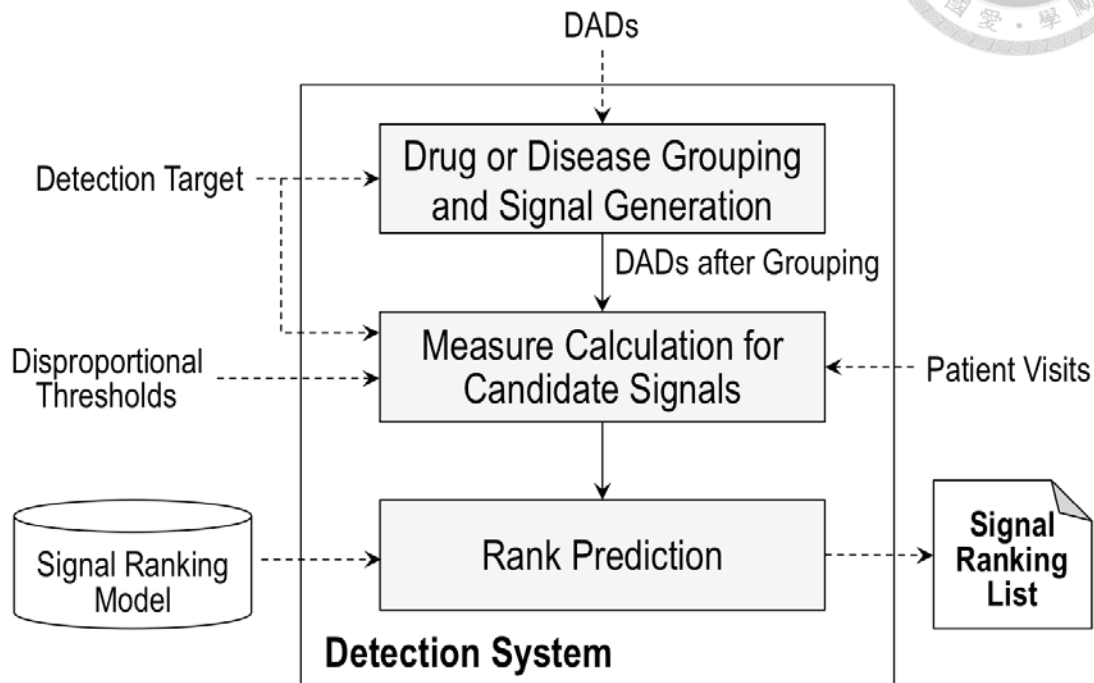
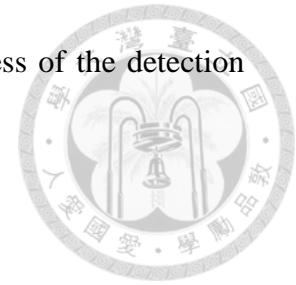


Figure 13: Detailed process of the detection system

3.4.1 Drug or Disease Grouping and Signal Generation

In this step, we map the ATC and ICD codes to the drug or disease group as indicated in the detection target, and then generate the drug-anchored or disease-anchored drug-outcome pairs as candidate signals. We generate the drug-anchored drug-outcome pairs by forming all the diseases relating to the drugs in DADs. Similar process can also be employed to generate disease-anchored drug-outcome pairs.

3.4.2 Measure Calculation for Candidate Signals

This step is similar to measure calculation for training data in the learning system.

Notice that the contingency table here is built by the same way of visit-based counting illustrated in Table 5. We then calculate all the measures mentioned in Chapter 3.3.3 for each drug-outcome pair on the basis of DADs as the testing data. We can also set some thresholds to prune insignificant pairs (e.g., the minimum number of drug-disease incidences is 10 and the ROR index is larger than 1.5). After these steps, the drug-outcome pairs are generated as candidate signals and used for rank prediction.

3.4.3 Rank Prediction

This step is the last part of our method, and we use the signal detection model to predict and rank all the candidate signals generated previously. The following is the function of how Ranking SVM determines the ranks of signals by the model constructed by training data set.

$$f(x; \hat{w}) = \langle \hat{w}, x \rangle$$

The function $f(x; \hat{w})$ gives every pair a score. By ordering these pairs by their scores, we can get the rank of all candidate signals.

Chapter 4 Evaluation and Results



In this chapter, we describe the experimental data, the design of our evaluation, and discuss our evaluation results.

4.1 Experimental Data

In this study, we used the Taiwan's National Health Insurance Research Database (NHIRD) covering the time-span of 2000-2009, and 1999 was also used for control window. The control window and the surveillance window we used were both 12 months.

Due to the high cost of query collection, we only choose four types of diseases as our disease-anchored detection targets. The reasons are listed as follows and Table 9 shows the detailed ICD codes for each disease type that we select.

1. Recently, many ADEs between new approval drugs and cardiovascular events are detected.
2. Cancer is a specific and serious side effect, and it is the disease that human beings want to prevent.
3. Humans metabolize drugs through liver or renal, which leaves some hepatotoxicity or acute renal toxicity on liver or renal affecting their functional act.

Table 9: Disease types and their corresponding ICD-9-CM codes

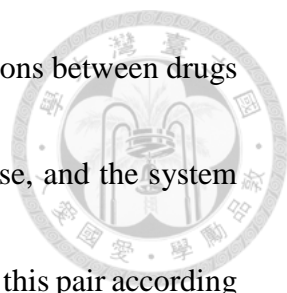
Disease Type	Corresponding ICD-9-CM
Cardiovascular events	402, 404, 410, 411, 413, 414, 424, 426, 427, 428, 7943
Hepatotoxicity	2774, 570, 573, 576, 7824
Cancer	140-208
Acute renal toxicity	584, 586

Several graduate students of School of Pharmacy at National Taiwan University and pharmacists in National Taiwan University Hospital were separated into two groups to label all disease-anchored drug-outcome pairs which were based on $a \geq 10$ and $ROR \geq 1.5$ in all types of diseases. Took acute renal toxicity as an example. We recruited eight labelers separated into two groups, and each group had four labelers. Every subset of disease-anchored drug-outcome pairs were labeled by two different pharmacists. Figure 14 shows the detailed information about labeling arrangement.

Acute Renal Toxicity Drug-ADR pair list		Expert	Experience	Work in
Subset 1		A	5 years and 4 months	Pharmacist of NTUH
		E	1 years and 9 months	Pharmacist of NTUH
Subset 2		B	5 years and 10 months	Pharmacist of NTUH
		F	4 years and 5 months	Pharmacist of NTUH
Subset 3		C	4 years and 4 months	Pharmacist of NTUH
		G	3 years and 8 months	Pharmacist of NTUH
Subset 4		D	9 months	Pharmacist of NTUH
		H	1 years and 5 months	Pharmacist of NTUH

Group 1 Group 2

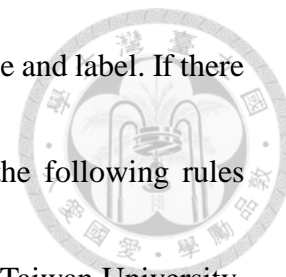
Figure 14: Labeler's arrangement and their work experiences



There is a system called Micromedex, which includes the relations between drugs and diseases. The labelers input the keyword of a drug and a disease, and the system will give this drug-outcome pair a definition. Then, labelers can label this pair according to the definition in Micromedex and their expert knowledge. For example, if they input Warfarin & Arrhythmia, and the system will return the answer “Atrial fibrillation - Thromboembolic disorder; Prophylaxis and FDA Labeled Indication.” Hence, this pair will be labeled as “Indication associations”.

The relation of “Indication association” could be clearly defined by this system. However, other relations might not be such definite and clear, so the labeler needs to make professional judgment when labeling. Take Rosiglitazone & Myocardial Infarction as another example. The Micromedex system gives the information as follow: “A meta-analysis of 52 double-blind, randomized, controlled clinical trials (mean duration, 6 months) showed a significantly increased incidence of myocardial infarction in patients who received rosiglitazone-containing therapy (n=10,039) compared with those who received alternative therapy, including placebo (n=6956; 0.4% vs 0.3%; odds ratio, 1.8; 95% CI, 1.03 to 3.25)....” According to the keywords of meta-analysis and significantly increased incidence, the labeler should attribute this pair to Known ADEs.

We believe that the experience of our labelers is well enough to judge and label. If there were pairs with inconsistent labels from two labelers, we used the following rules suggested by several professors of School of Pharmacy at National Taiwan University.



1. The “indication associations” need to be agreed by both two labelers. If labelers both label this pair as indication associations, this pair belongs to this label type. If they have inconsistent opinions, this pair will be removed from our dataset. If this pair is not labeled by both of the pharmacists as “indication associations,” go through the following steps.
2. If a drug-outcome pair is labeled by one or more of labeler as “Known ADEs,” it will be classified to this type. If not, go through the following steps.
3. If a drug-outcome pair is not classified as “Known ADEs” in the previous step and there is at least one pharmacist labeling it as “Suspected ADEs,” it will be classified to this type. If not, go through the following step.
4. A drug-outcome pair not included in the “Known ADEs” and “Suspected ADEs” will be labeled as “Unknown associations.”

The following table shows the number of all label types in the four disease (query) types we collected.

Table 10: Summary of our query and label collection

	Known ADRs	Suspected ADRs	Unknown Associations	Indication Associations
Cardiovascular events	376	43	1883	57
Cancer	24	5	300	43
Hepatotoxicity	346	176	585	44
Acute renal toxicity	230	87	768	21

4.2 Evaluation Design

As we mentioned in the literature review, the majority of EHR-based methods use single disproportional analysis measures. Thus, we employ RRR, PRR, ROR and BCPNN, four traditional and popular measures, to rank signals as benchmark of our study.

4.2.1 Evaluation Criteria

Evaluations on the performance of a ranking model are carried out by comparison between the ranking lists output by the model and the ranking lists given as ground truth. Several evaluation measures such as normalized discounted cumulative gain (NDCG), mean average precision (MAP), and winner take all (WTA) are widely used in information retrieval (IR) or other fields (Li, 2011).

There are two advantages of NDCG compared to many other measures. First, NDCG allows degrees of relevancy in each signal while most traditional ranking

measures only allow binary relevance. Second, NDCG involves a discount function over the rank while many other measures uniformly weight all positions (Wang, He, & Chen, 2013). Therefore, we take this measure as our evaluation criteria.



NDCG represents the normalized cumulative gain of accessing the information from position one to position k with discount on the positions. It is defined as

$$\text{NDCG}(k) = \text{DCG}_{max}^{-1}(k) \sum_{j:\pi_i(j) \leq k} G(j)D(\pi_i(j)),$$

where $\text{DCG}_{max}(k)$ is the normalizing factor and is chosen such that a perfect ranking π_i^* 's NDCG score at position k is 1. The gain function is normally defined as an exponential function of grade.

$$G(j) = 2^{y_{i,j}} - 1$$


where $y_{i,j}$ is the label of $d_{i,j}$ in ranking list π_i . The discount function is normally defined as a logarithmic function of position.

$$D(\pi_i(j)) = \frac{1}{\log_2(1 + \pi_i(j))}$$

where $\pi_i(j)$ is the position of $d_{i,j}$ in ranking list π_i . The DCG at position k for q_i becomes

$$\text{DCG}(k) = \sum_{j:\pi_i(j) \leq k} \frac{2^{y_{i,j}-1}}{\log_2(1 + \pi_i(j))}.$$

NDCG values are further average over queries ($i = 1, \dots, m$) (Li, 2011).



Perfect ranking	Formula	Explanation
(3, 3, 2, 2, 1, 1, 1)		grades: 3,2,1
(7, 7, 3, 3, 1, 1, 1)	$2^{y_{i,j}} - 1$	gains
(1, 0.63, 0.5, ...)	$1/\log(\pi_i(j) + 1)$	position discounts
(7, 11.41, 12.91, ...)	$\sum_{j:\pi_i(j)\leq k} \frac{2^{y_{i,j}} - 1}{\log(\pi_i(j) + 1)}$	DCG scores
(1/7, 1/11.41, 1/12.91, ...)	$DCG_{max}^{-1}(k)$	normalizing factors
(1,1,1,...)	$NDCG(k)$	NDCG scores
Imperfect ranking	Formula	Explanation
(2, 3, 2, 3, 1, 1, 1)		grades: 3,2,1
(3, 7, 3, 7, 1, 1, 1)	$2^{y_{i,j}} - 1$	gains
(1, 0.63, 0.5, ...)	$1/\log(\pi_i(j) + 1)$	position discounts
(3, 7.41, 8.91, ...)	$\sum_{j:\pi_i(j)\leq k} \frac{2^{y_{i,j}} - 1}{\log(\pi_i(j) + 1)}$	DCG scores
(1/7, 1/11.41, 1/12.91, ...)	$DCG_{max}^{-1}(k)$	normalizing factors
(0.43, 0.65, 0.69, ...)	$NDCG(k)$	NDCG scores

Figure 15: Example of NDCG (Li, 2011)

4.2.2 Evaluation Procedure

For each disease-anchored query, we randomly extracted 20% from each likelihood level of each disease query (the way we sampled called stratified sampling) as the training set, and the rest 80% of signals were used as the testing set. To improve the reliability of our evaluation, we performed thirty times of 20%-80% stratified random sampling for training and testing data. Thus, the evaluations of NDCG were the average of thirty random samples of four disease queries. The average training and testing sizes of four types of disease query are shown in Table 11.

Table 11: Average of training and testing set in each disease query

	Training set	Testing set
Cardiovascular events	472	1887
Cancer	75	297
Hepatotoxicity	231	920

Acute renal toxicity	223	883
----------------------	-----	-----



4.3 Comparative Evaluation

In this study, we try to figure out whether using multiple measures to rank the drug-outcome pairs would have better performance than single measure or not.

Table 12: Comparative evaluation results (NDCG@5 to NDCG@50)

	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@25	NDCG@30	NDCG@35	NDCG@40	NDCG@45	NDCG@50
PRR	0.229601	0.292851	0.300054	0.313593	0.337741	0.351985	0.360424	0.36322	0.374212	0.383233
ROR	0.224791	0.29325	0.29829	0.311864	0.3372	0.351699	0.361749	0.369087	0.37582	0.382447
RRR	0.227669	0.293766	0.302943	0.313998	0.339622	0.354806	0.361625	0.36503	0.374308	0.382599
BCPNN	0.256174	0.290166	0.306456	0.323195	0.335955	0.346342	0.354974	0.364896	0.378762	0.386311
Ranking SVM (All Measures)	0.375349	0.367747	0.375066	0.383424	0.387503	0.393022	0.400416	0.404632	0.405939	0.410429

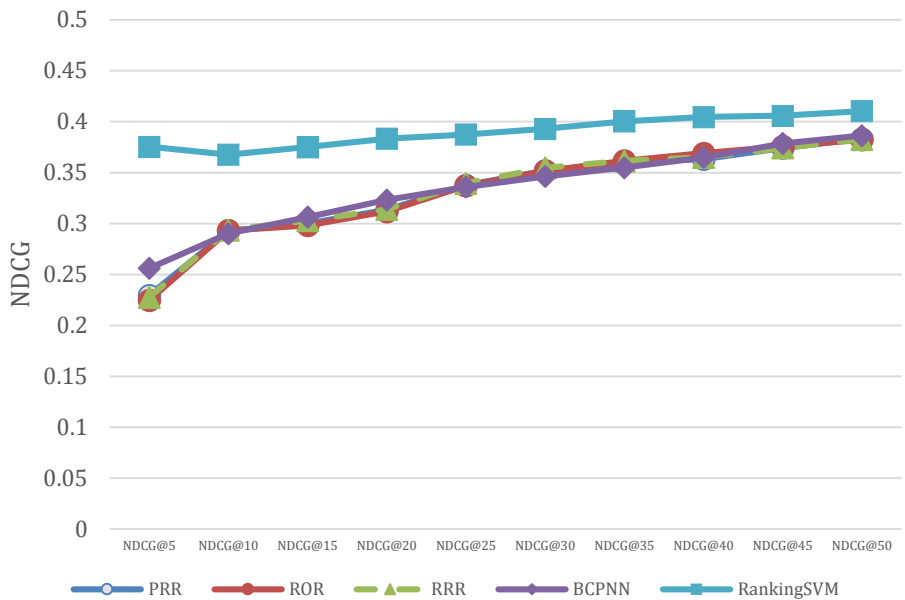


Figure 16: Comparative evaluation results (NDCG@5 to NDCG@50)

As Table 12 and Figure 16 show, our proposed ranking method (using the learning to rank approach with all measures) outperforms all benchmarks (i.e., single measure,

such as PRR, ROR, RRR, and BCPNN).



4.4 Additional Evaluations

In this section, we design four additional experiments. First, we want to know the effect of variables selection. Second, we examine the effect of training size on the detection effectiveness of our proposed ranking method. Third, we investigate whether different control and surveillance window sizes will affect the effectiveness of our proposed ranking system. Fourth, we attempt to evaluate the appropriateness of non-mono-domain training.

4.4.1 Experiment 1: Effects of Variables Selection

In this experiment, we want to know whether different types of measures influence the effectiveness in Ranking SVM. Thus, we design four types of datasets with different subsets of measures for ranking.

1. Type 1: Four traditional disproportional analysis measures, including RRR, PRR, ROR and BCPNN.
2. Type 2: Traditional measures plus other association rule measures introduced in Table 4.
3. Type 3: Traditional measures plus LDA measures.

4. Type 4: Traditional measures plus other association rule measures and LDA measures.

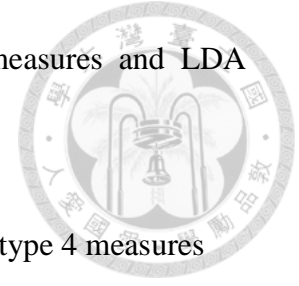


Table 13: NDCG evaluation for using type 1, type 2, type 3 and type 4 measures (using Ranking SVM)

	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@25	NDCG@30	NDCG@35	NDCG@40	NDCG@45	NDCG@50
Type 1	0.35014	0.368536	0.373013	0.378923	0.382973	0.389782	0.395184	0.404674	0.415084	0.422082
Type 2	0.333855	0.338281	0.356462	0.365808	0.369546	0.379703	0.388721	0.394187	0.3962	0.401202
Type 3	0.380331	0.414653	0.423439	0.434017	0.444889	0.451511	0.45446	0.459503	0.463696	0.466767
Type 4	0.375349	0.367747	0.375066	0.383424	0.387503	0.393022	0.400416	0.404632	0.405939	0.410429

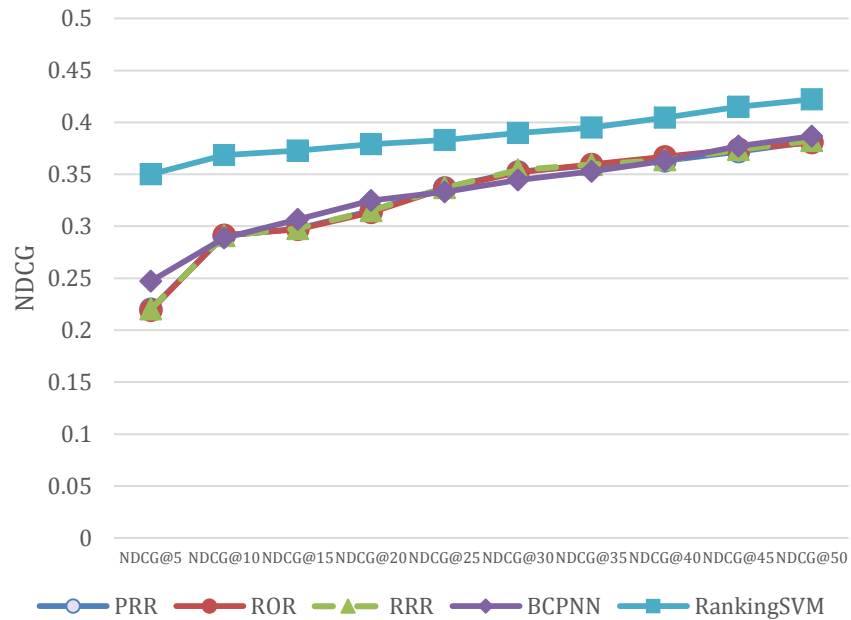


Figure 17: NDCG evaluation on type 1

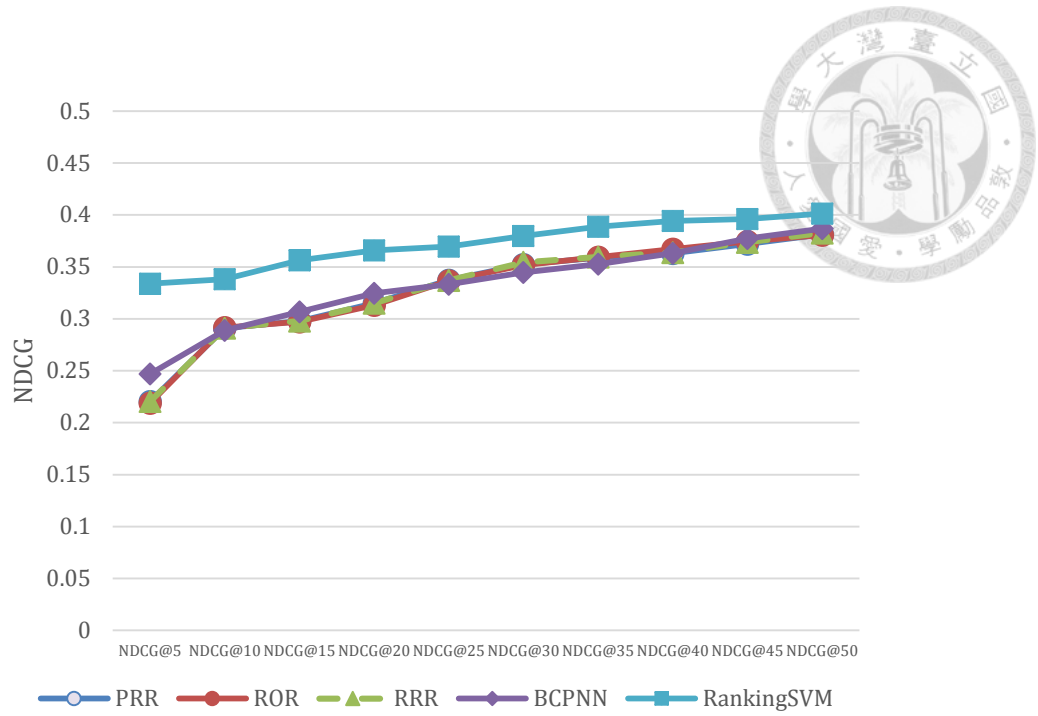


Figure 18: NDCG evaluation on type 2

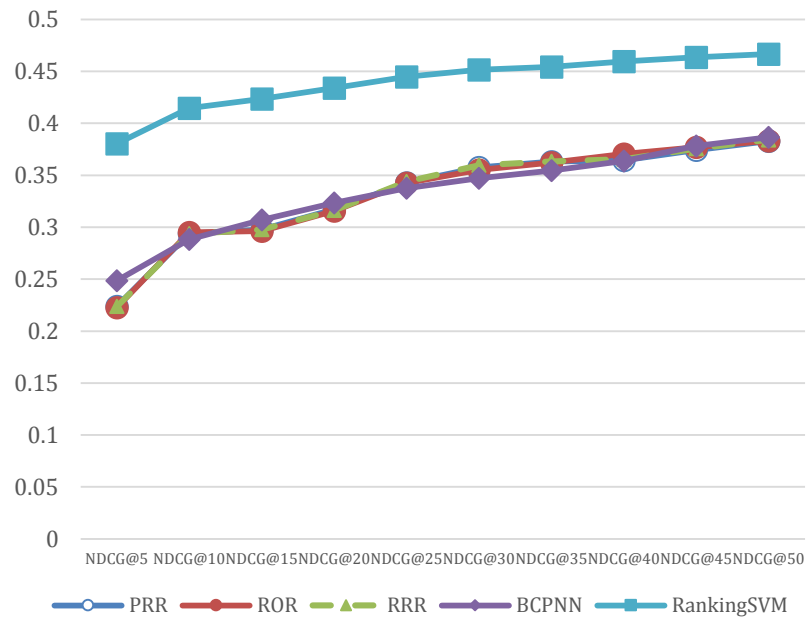


Figure 19: NDCG evaluation on type 3

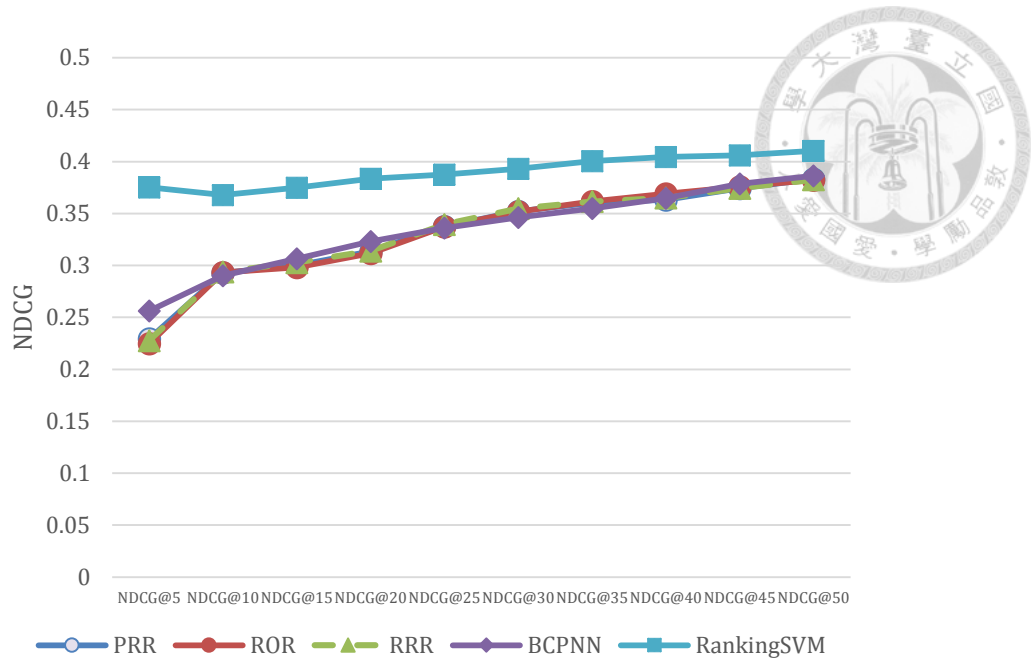


Figure 20: NDCG evaluation on type 4

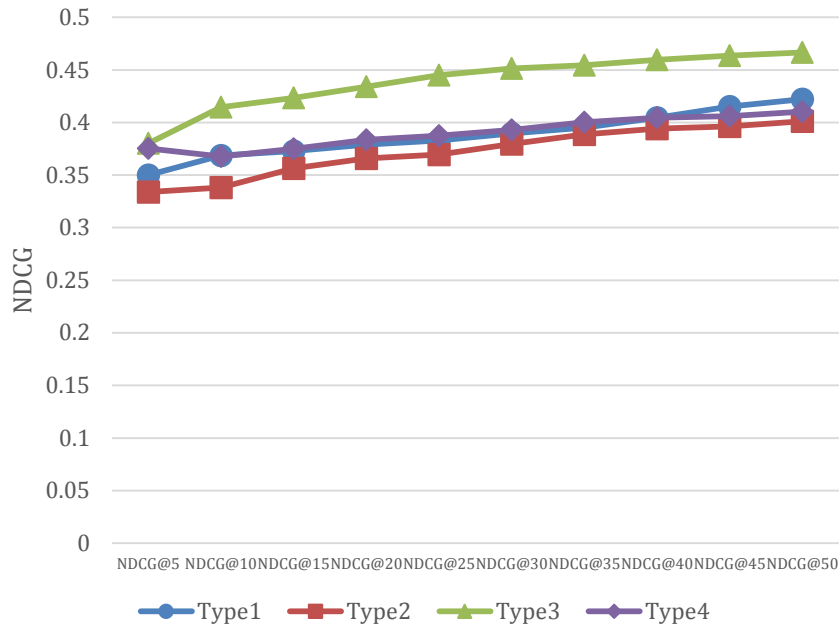
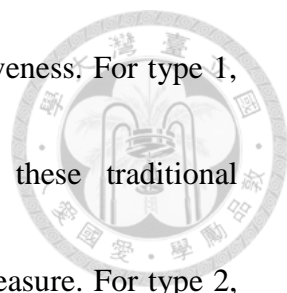


Figure 21: NDCG evaluation across four types of measures

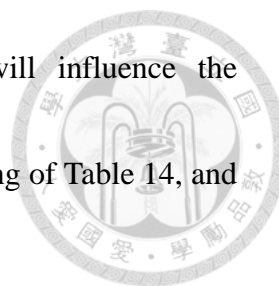
From Table 13 and Figures 17-20, we can observe that the performance of our proposed ranking method using different types of measures are higher than that of the benchmarks in NDCG evaluation. Moreover, we can also observe that the incorporation



of other association rule measures deteriorate the detection effectiveness. For type 1, the learning to rank method learns the ranking ability of these traditional disproportional analysis measures and outperforms every single measure. For type 2, we add other association rule measures, and the performance rushes down to the poorest, among all types of measures examined. We surmise that other association rule measures provide less useful information for ranking than traditional disproportional analysis measures. For type 3, it seems that the LDA method can extract some implicit relations between drugs and diseases, and the effectiveness achieved by this type noticeably outperforms that of type 1 or type 2. For type 4, because the benefits of incorporating LDA measures are offset by the involvement of other association rule measures, its effectiveness (of type 4) is worse than that of type 3, but better than that of type 2.

4.4.2 Experiment 2: Effects of Training Sizes

In this experiment, we want to know whether we can decrease training size and how much we can decrease without sacrificing too much effectiveness. Thus, we decrease the training size from 20% to 5% (in decrements of 5%) by stratified random sampling. We use the same testing size of 80%, because different testing sizes will cause the evaluation of NDCG unfair. The smaller the testing size, the larger the NDCG value.



Tables 14 and 15 demonstrate how different testing sizes will influence the corresponding NDCG values. Table 15 is the 50% stratified sampling of Table 14, and the NDCG of Table 15 is higher than Table 14 at top 5.

Table 14: Example explaining the effect of testing size on NDCG (all ranks)

Perfect rank	4	4	4	4	3	3	2	1
Gains	15	15	15	15	7	7	3	1
Position discount	1	0.630929754	0.5	0.430676558	0.386852807	0.356207187	0.333333333	0.315464877
DCG	15	24.4639463	31.9639463	38.42409467	41.13206433	43.62551464	44.62551464	44.94097951
NDCG	0.066666667	0.04087648	0.031285248	0.026025337	0.024311933	0.022922366	0.022408705	0.022251406
Real rank	3	4	2	4	1	4	4	3
Gains	7	15	3	15	1	15	15	7
Position discount	1	0.630929754	0.5	0.430676558	0.386852807	0.356207187	0.333333333	0.315464877
DCG	7	16.4639463	17.9639463	24.42409467	24.81094748	30.15405529	35.15405529	37.36230943
NDCG	<u>0.066666667</u>	<u>0.04087648</u>	<u>0.031285248</u>	<u>0.026025337</u>	<u>0.024311933</u>	0.022922366	0.022408705	0.022251406

Table 15: Example explaining the effect of testing size on NDCG (50% stratified sampling of all ranks)

Perfect rank	4	4	3	2	1
Gains	15	15	7	3	1
Position discount	1	0.630929754	0.5	0.430676558	0.386852807
DCG	15	24.4639463	27.9639463	29.25597598	29.64282879
NDCG	0.066666667	0.04087648	0.035760332	0.034181051	0.033734972
Real rank	3	4	2	4	1
Gains	7	15	3	15	1
Position discount	1	0.630929754	0.5	0.430676558	0.386852807
DCG	7	16.4639463	17.9639463	24.42409467	24.81094748
NDCG	<u>0.066666667</u>	<u>0.04087648</u>	<u>0.035760332</u>	<u>0.034181051</u>	<u>0.033734972</u>

From Table 16 and Figures 22 to 26, we can figure out the decrease of training size drop down the performance a little, but all different training sizes outperform the benchmarks. That is, small training size (e.g., 5%) still has the ability to predict signal ranks. Therefore, from the practical consideration, labeling 5% of signals as the training set may not be a difficult job and our proposed ranking method (using the learning to rank approach) is practically viable and appealing.

Table 16: NDCG evaluation for different training sizes

	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@25	NDCG@30	NDCG@35	NDCG@40	NDCG@45	NDCG@50
PRR	0.240932	0.364139	0.33922	0.355694	0.390862	0.41761	0.428886	0.424895	0.444472	0.455363
ROR	0.238208	0.367709	0.337822	0.353746	0.391924	0.415405	0.430418	0.435565	0.44777	0.454743
RRR	0.240932	0.364139	0.33922	0.355732	0.39192	0.421651	0.429799	0.426309	0.445616	0.457734
BCPNN	0.240932	0.364139	0.33922	0.355732	0.39192	0.421651	0.429799	0.426309	0.445616	0.457734
Train-20%	0.375349	0.367747	0.375066	0.383424	0.387503	0.393022	0.400416	0.404632	0.405939	0.410429
Train-15%	0.371966	0.374065	0.380272	0.385796	0.390473	0.39763	0.404003	0.407734	0.408679	0.41221
Train-10%	0.363071	0.369609	0.376426	0.373986	0.378987	0.382819	0.390417	0.394565	0.400337	0.402749
Train-5%	0.367237	0.361576	0.370987	0.371427	0.377014	0.383246	0.388726	0.391149	0.394567	0.397769

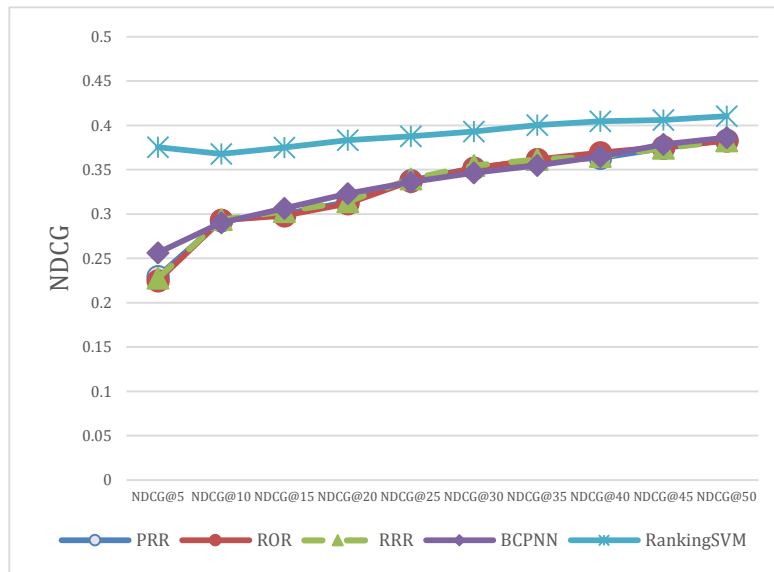


Figure 22: NDCG evaluation for 20% training size

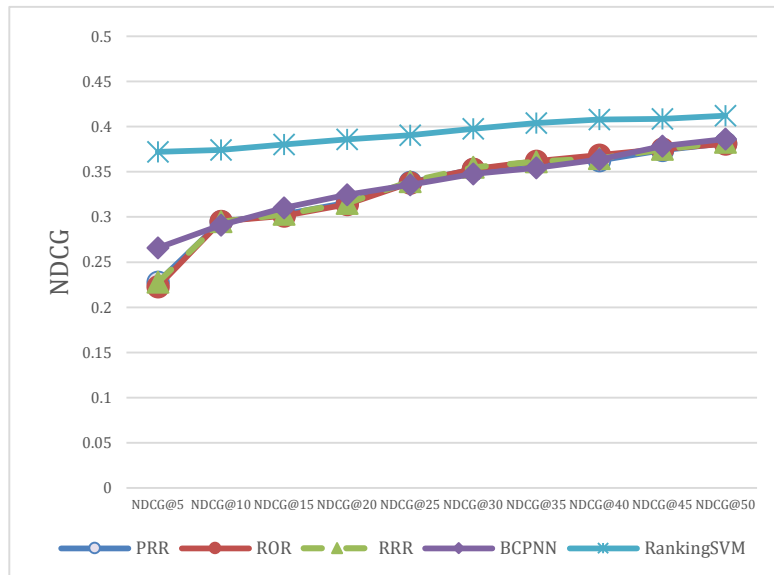


Figure 23: NDCG evaluation for 15% training size

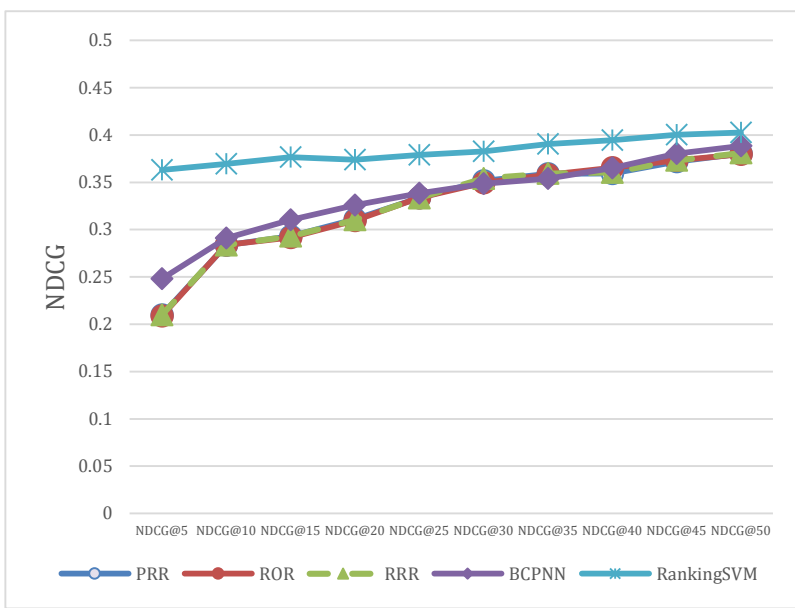


Figure 24: NDCG evaluation for 10% training size

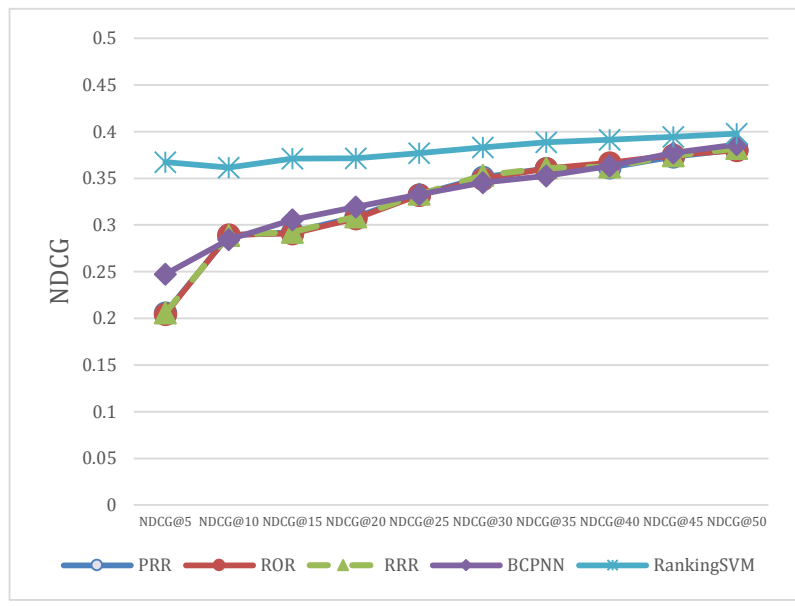


Figure 25: NDCG evaluation for 5% training size

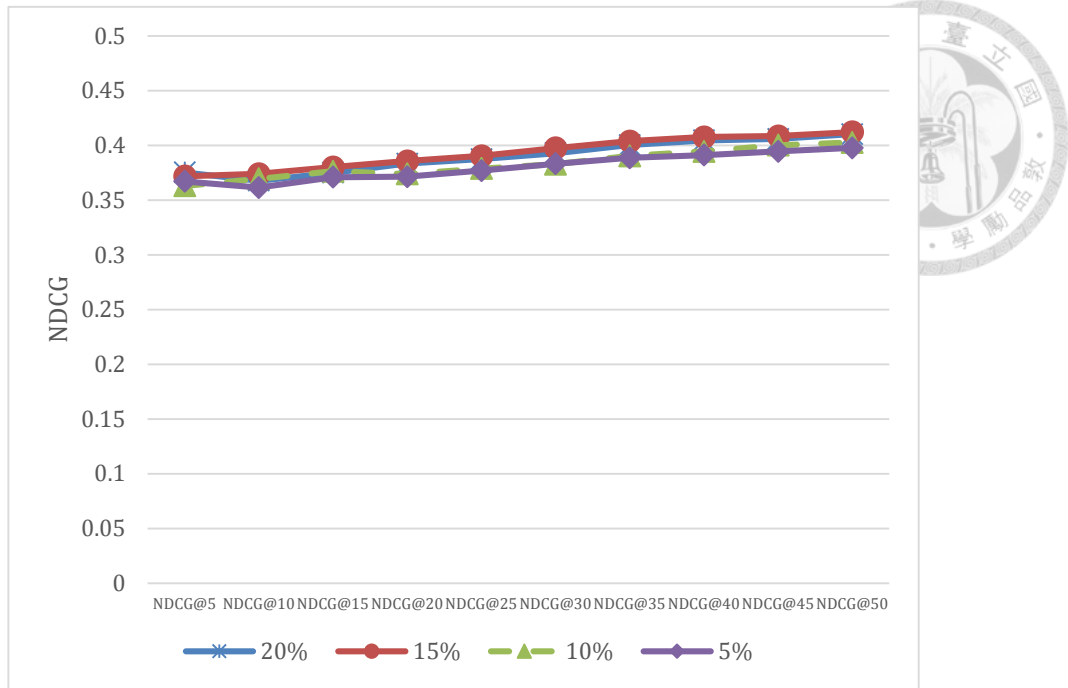


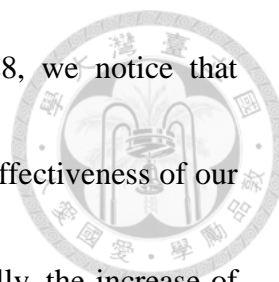
Figure 26: NDCG evaluation across different training sizes (using Ranking SVM)

4.4.3 Experiment 3: Effects of Surveillance and Control Window Sizes

In this experiment, we attempt to examine the effects of sizes of control and surveillance windows on the detection effectiveness of our proposed ranking method.

We design two experiments that vary the size of control window and that of the surveillance window independently. For surveillance window, we fix the surveillance window as 12 months and examine the sizes of control window from 1, 3, 6 to 12 months. For control window, we fix the control window as 12 months and vary the sizes of surveillance window from 1, 3, 6, to 12 months.

From Table 17 and Figure 27, we observe that the change of control window size does not have significant influence on the detection effectiveness of our proposed



ranking method. However, according to Table 17 and Figure 28, we notice that surveillance window sizes have stronger impacts on the detection effectiveness of our proposed ranking method than control window sizes do. Specifically, the increase of surveillance window size from 1 month to 12 months generally improves detection effectiveness. Thus, we perform a detailed look into each disease-anchored query of different surveillance window sizes. Before comparing the performance of each query, we show the change of drug-outcome pairs that can be detected under different surveillance window sizes in Table 18. When we decrease the surveillance window size from 12 months to 1 month, the number of drug-outcome pairs associated to hepatotoxicity and acute renal toxicity decreases with a greater magnitude than that associated to cardiovascular events and cancer.

Table 17: Effects of sizes of control window and surveillance window (where c12_s12 means that control window of 12 months and surveillance window of 12 months)

	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@25	NDCG@30	NDCG@35	NDCG@40	NDCG@45	NDCG@50
c12_s12	0.375349	0.367747	0.375066	0.383424	0.387503	0.393022	0.400416	0.404632	0.405939	0.410429
c12_s6	0.340878	0.333196	0.344085	0.355006	0.36966	0.375812	0.376511	0.381115	0.386412	0.395126
c12_s3	0.327075	0.327599	0.32805	0.333469	0.345429	0.355695	0.36823	0.375882	0.382686	0.388666
c12_s1	0.277644	0.312992	0.327247	0.328716	0.338755	0.347565	0.355814	0.363879	0.371259	0.37847
c6_s12	0.322468	0.342734	0.36779	0.373795	0.384859	0.392265	0.398971	0.404367	0.40828	0.411226
c3_s12	0.320167	0.339235	0.35998	0.36911	0.378844	0.390089	0.39451	0.400711	0.40513	0.41034
c1_s12	0.336869	0.348018	0.360918	0.371872	0.380472	0.389986	0.39258	0.398455	0.403156	0.40813

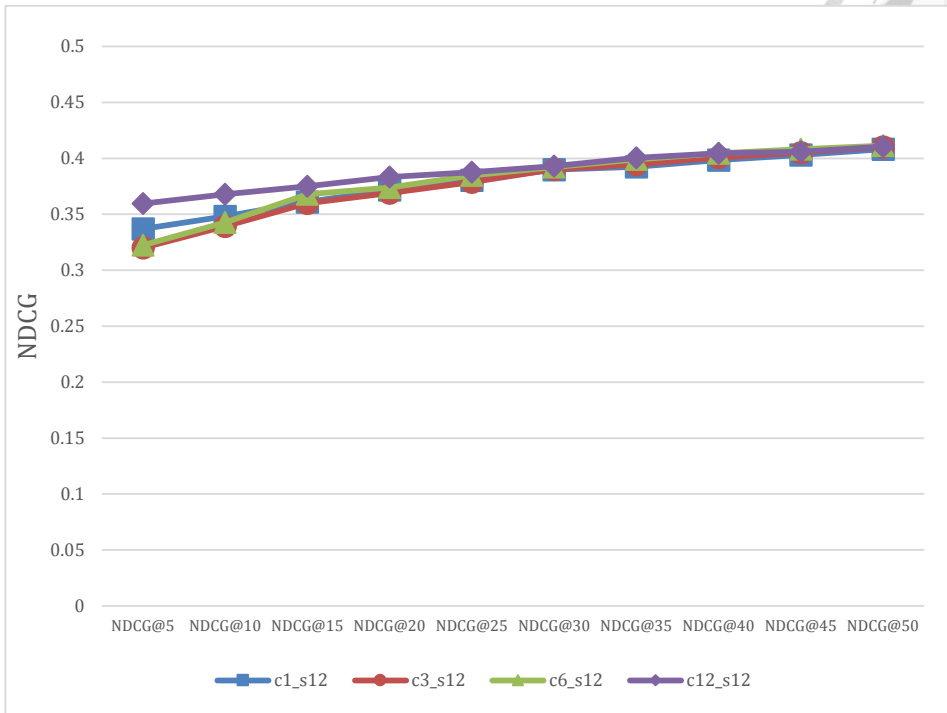
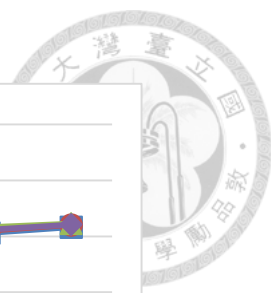


Figure 27: NDCG evaluation of different control window sizes

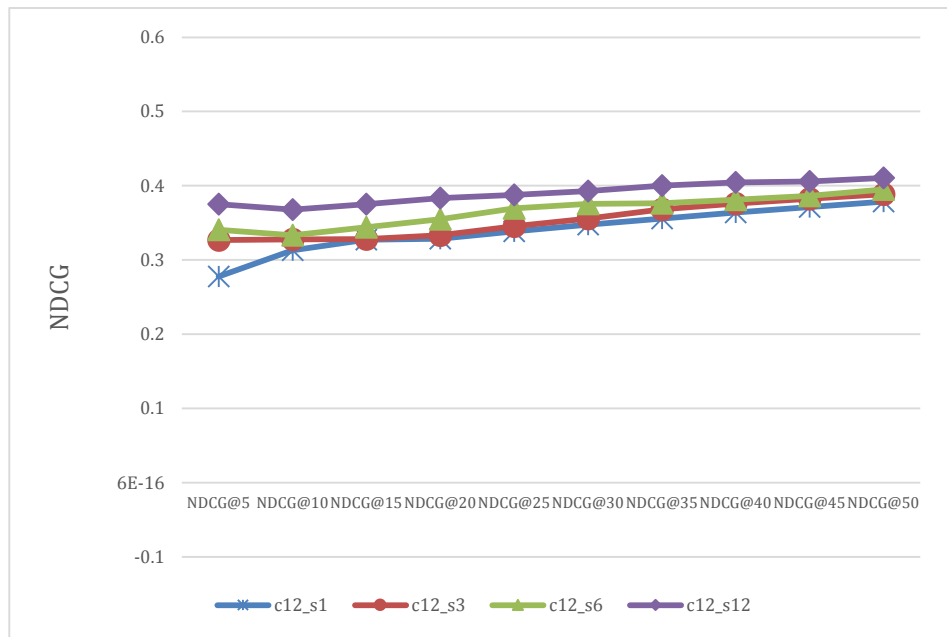


Figure 28: NDCG evaluation of different surveillance window sizes

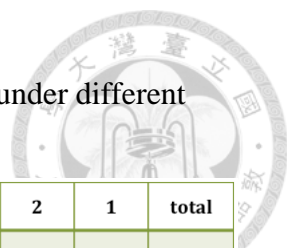


Table 18: Number of drug-outcome pairs in each disease query under different surveillance window sizes

c12s12	4	3	2	1	total
Cardiovascular events	376	43	1883	57	2359
Cancer	24	5	300	43	372
Hepatotoxicity	346	176	585	44	1151
Acute renal toxicity	230	87	768	21	1106

c12s6	4	3	2	1	total
Cardiovascular events	376	43	1883	57	2359
Cancer	24	5	300	43	372
Hepatotoxicity	343	173	579	43	1138
Acute renal toxicity	221	86	748	21	1076

c12s3	4	3	2	1	total
Cardiovascular events	375	43	1883	57	2358
Cancer	24	5	300	43	372
Hepatotoxicity	336	171	569	43	1119
Acute renal toxicity	220	84	731	20	1055

c12s1	4	3	2	1	total
Cardiovascular events	370	43	1871	57	2341
Cancer	24	5	298	43	370
Hepatotoxicity	320	164	545	41	1070
Acute renal toxicity	204	81	689	19	993

Figures 29 to 32 illustrates the effects of surveillance window sizes on the detection effectiveness for each disease-anchored query. It is obviously that detecting cancer events (ADRs) needs longer surveillance window. Cardiovascular events may include both short-term and long-term, so that 3 and 12 months of surveillance window perform better. In this experiment, we only explain what we have observed, and more thorough analyses require inputs and insights from domain experts.

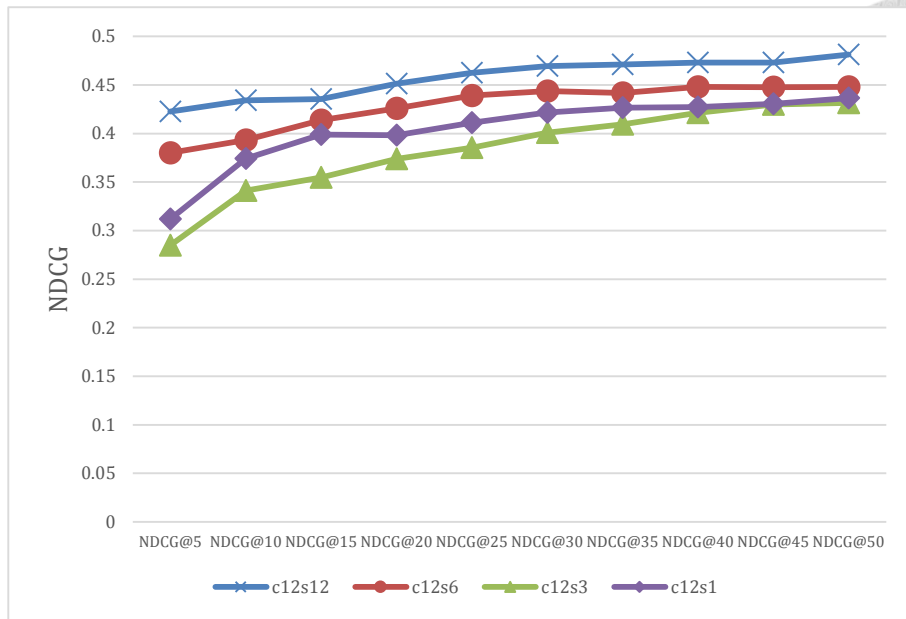


Figure 29: NDCG evaluation of different surveillance window sizes (for hepatotoxicity)

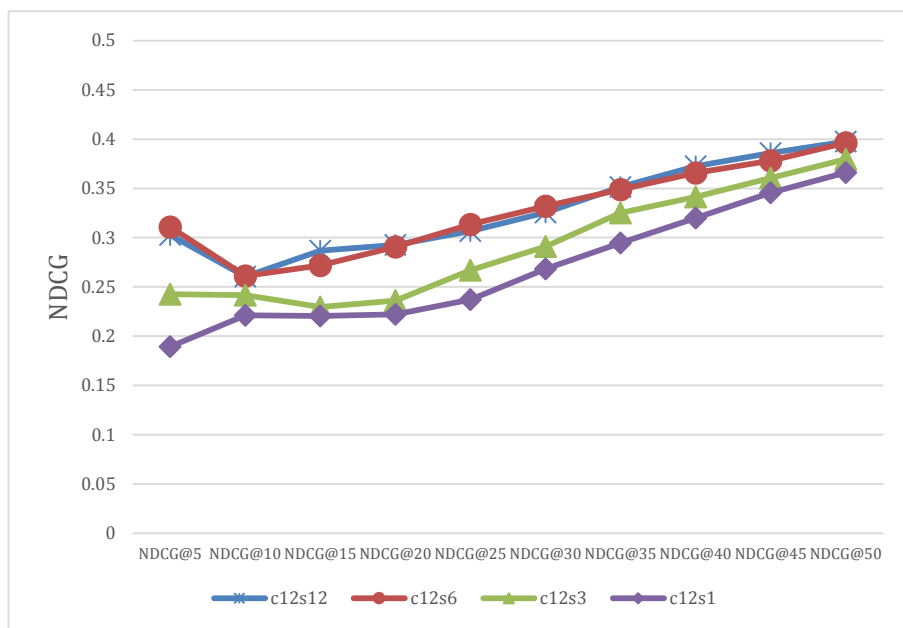


Figure 30: NDCG evaluation of different surveillance window sizes (for cancer)

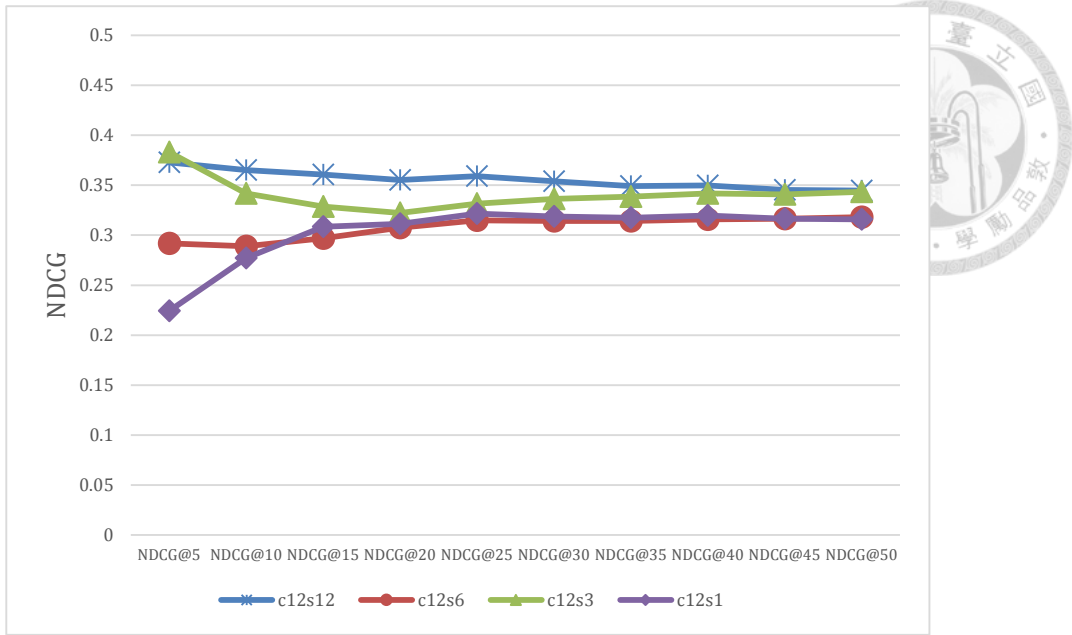


Figure 31: NDCG evaluation of different surveillance window sizes (for cardiovascular events)

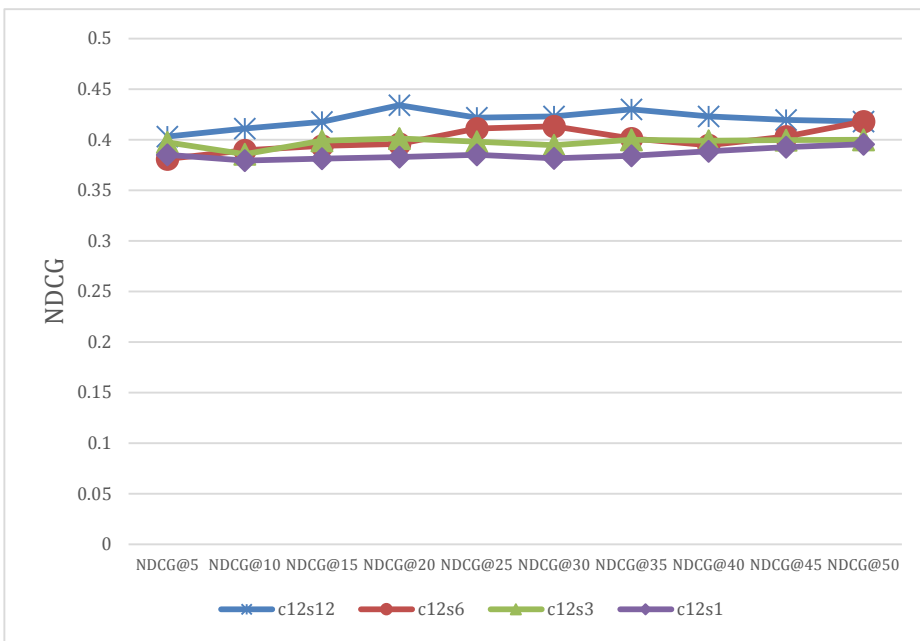


Figure 32: NDCG evaluation of different surveillance window sizes (for acute renal toxicity)

4.4.4 Experiment 4: Appropriateness of Non-Mono-Domain Training

In our previous experiments and evaluations, we use mono-domain training. That is, for each disease-anchored query, we randomly select a certain percentage of drug-outcome pairs from the query for training purpose and use the remaining drug-outcome pairs from the same query for testing purpose. In this experiment, we attempt to examine the feasibility of non-mono-domain training. Specifically, we design two non-mono-domain training scenarios: cross-domain training and mixed domain training.

Figure 33 illustrates the scenario of cross-domain training. Specifically, we use the drug-outcome pairs from some queries for training purpose and use the drug-outcome pairs from the remaining query for testing purpose. For example, as mentioned, we have 4 disease-anchored queries (i.e., cardiovascular events, cancer, hepatotoxicity, and acute renal toxicity). Thus, in the cross-domain training scenario, we may use the drug-outcome pairs from three queries (e.g., cardiovascular events, cancer, and hepatotoxicity) as the training set and employ the drug-outcome pairs from the remaining one query (i.e., acute renal toxicity) as the testing set.

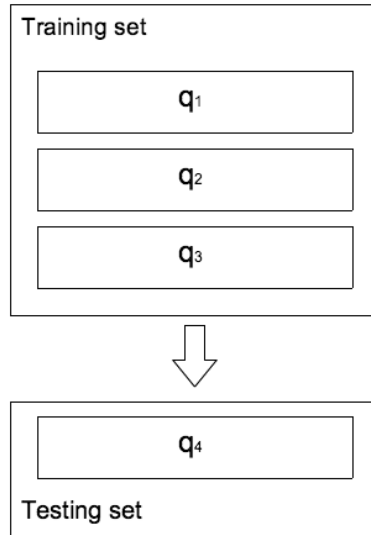


Figure 33: Illustration of cross-domain training

From Table 19 and Figure 34, we can find that this cross-domain training scenario achieves worse detection effectiveness than some benchmarks. Such inferior performance may be caused by that the characteristics of ADEs of different disease queries may be different. Although we combine three disease queries for training purpose, their characteristics may be different from those of the testing query, which weakening the detection effectiveness of our proposed ranking method.

Table 19: NDCG evaluations for the cross-domain training scenario

	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@25	NDCG@30	NDCG@35	NDCG@40	NDCG@45	NDCG@50
PRR	0.222721	0.285462	0.297434	0.311365	0.336339	0.352086	0.360764	0.363177	0.376467	0.38452
ROR	0.222721	0.28555	0.296466	0.309945	0.336611	0.351388	0.36136	0.369192	0.377857	0.384207
RRR	0.244614	0.323644	0.332241	0.338287	0.354039	0.361794	0.364732	0.370055	0.387492	0.399487
BCPNN	0.251998	0.291074	0.310819	0.326193	0.341293	0.34977	0.356004	0.367305	0.382263	0.390585
Ranking SVM - Cross-Domain	0.262384	0.282681	0.30854	0.317723	0.325441	0.335989	0.347397	0.354729	0.360503	0.36958

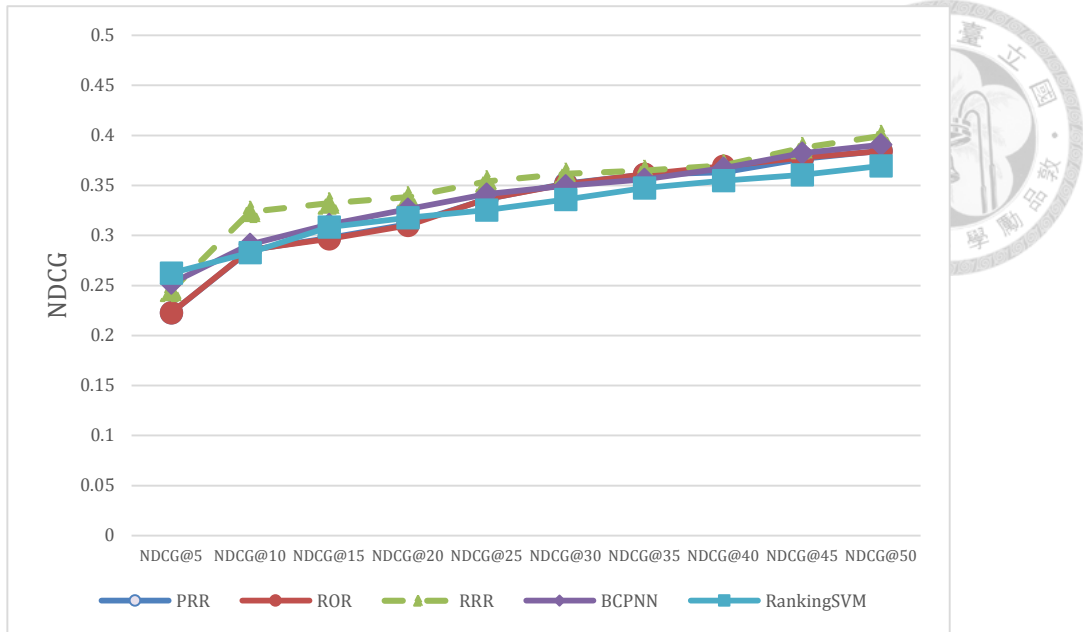


Figure 34: NDCG evaluation for cross-domain training

The second scenario for non-mono-domain training is the mixed domain. As illustrated in Figure 35, we use part of drug-outcome pairs from all disease queries to form a training set for building a signal ranking model, and then use the remaining drug-outcome pairs of each disease query for testing purpose.

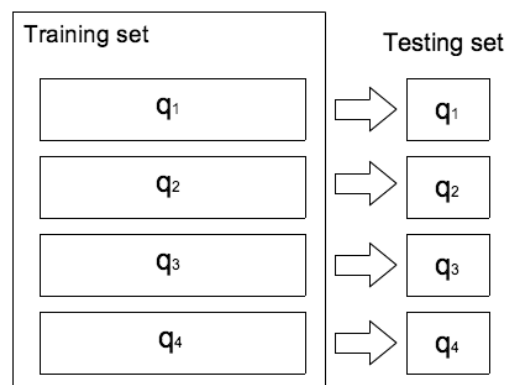


Figure 35: Illustration of mixed-domain training

From Table 20 and Figure 36, we can observe that the performance of the mixed-domain training does not go beyond the benchmarks. Although the training set contain

drug-outcome pairs from the same disease query as the testing set, the training set also include drug-outcome pairs from other disease queries, which undermines detection effectiveness. The detection effectiveness of the mixed-domain training scenario appears to be better than that of the cross-domain training scenario. This result suggests the utility of the inclusion of drug-outcome pairs (into the training set) from the same disease query as the testing set.

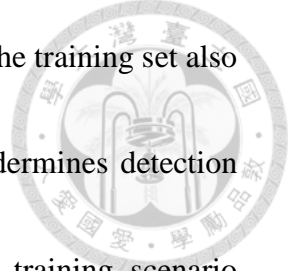


Table 20: NDCG evaluations for mixed-domain training

	NDCG@5	NDCG@10	NDCG@15	NDCG@20	NDCG@25	NDCG@30	NDCG@35	NDCG@40	NDCG@45	NDCG@50
PRR	0.323043	0.372997	0.402165	0.422755	0.43396	0.443417	0.449631	0.457553	0.46569	0.475247
ROR	0.321797	0.37059	0.401533	0.423635	0.433261	0.443882	0.449794	0.458949	0.465195	0.475084
RRR	0.32195	0.373946	0.40259	0.423266	0.434229	0.443879	0.449073	0.45816	0.466334	0.475348
BCPNN	0.336302	0.38266	0.412145	0.430298	0.43754	0.442887	0.451806	0.46152	0.470519	0.478853
Ranking SVM - Mixed Domain	0.300185	0.35885	0.388277	0.409716	0.423674	0.436056	0.444196	0.454677	0.461726	0.470559

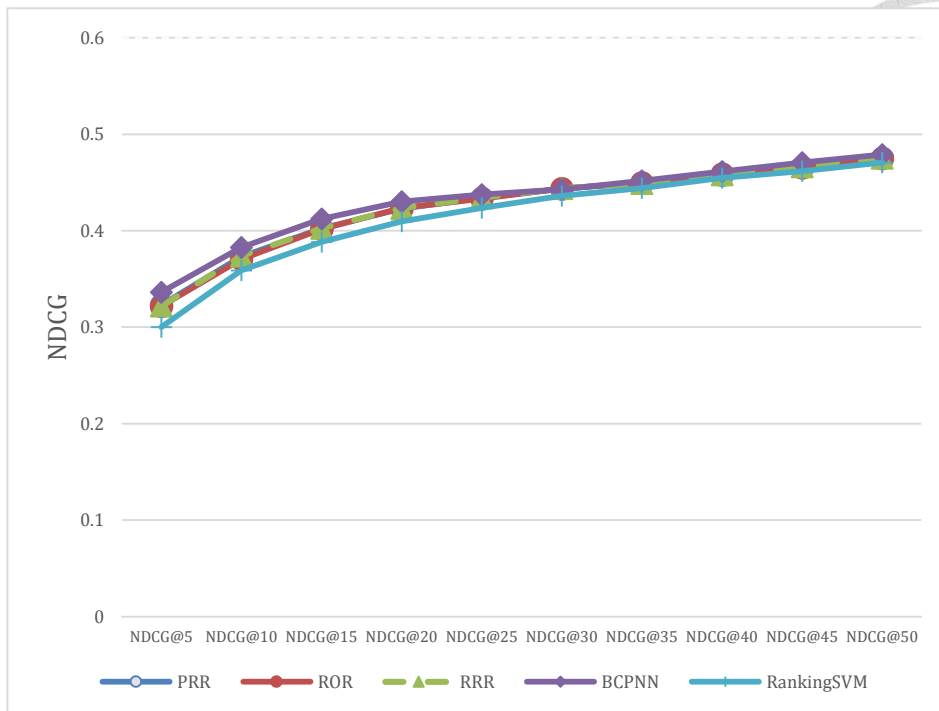
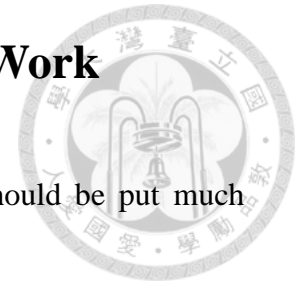


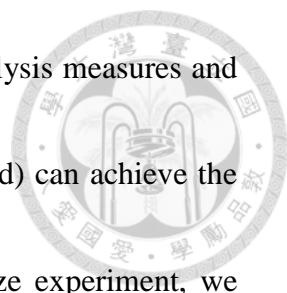
Figure 36: NDCG evaluation for mixed-domain training

Chapter 5 Conclusion and Future Work




Pharmacovigilance is a serious and worldwide issue that should be put much concern on it. Many drug safety signal detection methods for SRS databases have been developed in the literature. Due to the problems of volunteer nature in SRS-based methods, researchers start to investigate and develop EHR-based drug safety signal detection methods. Extended from SRS-based methods, EHR-based methods generally use single disproportional analysis measure, which ranks candidate drug-outcome pairs by the value of the selected measure.

In this study, we develop a supervised learning method (i.e., learning to rank) that learns from a collection of ranked drug-outcome pairs a signal ranking model, which can be employed to rank for unranked candidate drug-outcome pairs. In addition, we try to extract implicit relations between drugs and diseases by using the LDA method and develop additional measures for ranking purpose. Our empirical evaluation results suggest that our proposed ranking method (using Ranking SVM as the underlying learning to rank method) significantly outperforms the benchmarks (i.e., using single disproportional analysis measures). In addition, we conduct four additional experiments that provide more in-depth analyses. First, in the effects of variable selection



experiment, we find that combining traditional disproportional analysis measures and drug-disease association measures (extracted from the LDA method) can achieve the best detection effectiveness. Second, in the effects of training size experiment, we observe that smaller training size only slightly lower the detection effectiveness of our proposed ranking method, and the smallest training size (i.e., 5%) examined in this experiment still performs better than benchmarks. Third, in the effects of surveillance and control window size experiment, the control window size does not influence much on the performance of ranking, but the surveillance window size seems to have greater influence. We also observe that different disease queries may require different surveillance window sizes. Finally, in the non-mono-domain training experiment, we show that cross-domain and mixed-domain training scenarios perform worse than the benchmarks. The mono-domain training still represents the best design.

There are some limitations and further research directions relevant to our study. First, we suggest expanding the number of disease-anchored queries in order to improve the reliability of our evaluation results. In our study, we only have four disease queries (i.e., cardiovascular events, cancer, hepatotoxicity, and acute renal toxicity). In the future, we should collect more disease queries and associated labeled drug-outcome



pairs for evaluation purposes. Second, incorporating more measures that are potentially useful to drug safety signal detection in the claims databases may further improve the effectiveness of our proposed ranking method. Third, there exist further research directions for drug safety signal detection in EHR. For example, it would be essential to detect ADEs from EHR databases with the consideration of patients' preexisting medical conditions. In addition, it is also imperative to develop appropriate detection methods capable of detecting drug interactions and dose-related ADEs from EHR databases.

References



- (US), N. C. for H. S. (1980). ICD-9-CM: International Classification of Diseases 9th Revision Clinical Modification. *US Department of Health and Human Services, Public Health Service, Health Care Financing Administration.*
- Almenoff, J. S., Pattishall, E. N., Gibbs, T. G., DuMouchel, W., Evans, S. J. W., & Yuen, N. (2007). Novel statistical tools for monitoring the safety of marketed drugs. *Clinical Pharmacology and Therapeutics*, 82(2), 157–66.
- Azevedo, P. J., & Jorge, M. (2007). Comparing Rule Measures for Predictive Association Rules. In *Machine Learning: ECML 2007* (pp. 510–517). Springer Berlin Heidelberg.
- Balakin, K. (2009). *Pharmaceutical data mining: approaches and applications for drug discovery*. John Wiley & Sons.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Cao, Y., Xu, J., Liu, T., Li, H., Huang, Y., & Hon, H. (2006). Adapting Ranking SVM to Document Retrieval. *ACM SIGIR Conference*, (49), 186–193.
- Cao, Z., Qin, T., Liu, T., Tsai, M., & Li, H. (2007). Learning to rank: from pairwise approach to listwise approach. *Proceedings of the 24th International Conference on Machine Learning*.
- Chapelle, O., & Keerthi, S. S. (2009). Efficient Algorithms for Ranking with SVMs. *Information Retrieval Journal*.
- Chen, L., Zeng, W.-M., Cai, Y.-D., Feng, K.-Y., & Chou, K.-C. (2012). Predicting Anatomical Therapeutic Chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. *PloS One*, 7(4), e35254.
- Choi, N., Chang, Y., Kim, J., Choi, Y., & Park, B. (2011). Comparison and validation of data-mining indices for signal detection : using the Korean national health

insurance claims database. *Pharmacoepidemiology and Drug Safety*, 20, 1278–1286.

Coloma, P. M., Trifirò, G., Patadia, V., & Sturkenboom, M. (2013). Postmarketing safety surveillance : where does signal detection using electronic healthcare records fit into the big picture? *Drug Safety : An International Journal of Medical Toxicology and Drug Experience*, 36(3), 183–97.

Drug, N. (n.d.). Standard Operating Procedure.

[Http://pvtoolkit.org/toolkit/readers/sop_spontaneous_reporting.doc](http://pvtoolkit.org/toolkit/readers/sop_spontaneous_reporting.doc). Retrieved from http://pvtoolkit.org/toolkit/readers/sop_spontaneous_reporting.doc

Evans, S. J., Waller, P. C., & Davis, S. (2001). Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports. *Pharmacoepidemiology and Drug Safety*, 10(6), 483–486.

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

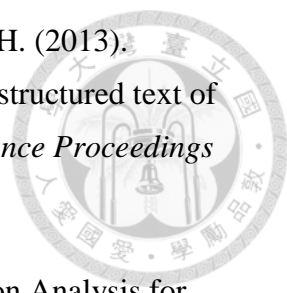
Harpaz, R., DuMouchel, W., Shah, N. H., Madigan, D., Ryan, P., & Friedman, C. (2012). Novel data-mining methodologies for adverse drug event discovery and analysis. *Clinical Pharmacology and Therapeutics*, 91(6), 1010–1021.

Hauben, M., & Bate, A. (2009). Decision support methods for the detection of adverse events in post-marketing data. *Drug Discovery Today*, 14(7-8), 343–357.

Hazell, L., & Shakir, S. A. W. (2006). Under-Reporting of Adverse A Systematic Review. *Drug Safety : An International Journal of Medical Toxicology and Drug Experience*, 29(5), 385–396.

Health, I. N. P. (2006). *The SAFETY of MEDICINES IN PUBLIC HEALTH PROGRAMMES: Pharmacovigilance an essential tool*. World Health Organization.

Hoerbst, a, & Ammenwerth, E. (2010). Electronic health records. A systematic review on quality requirements. *Methods of Information in Medicine*, 49(4), 320–336.

- 
- Iyer, S. V, Lependu, P., Harpaz, R., Bauer-Mehren, A., & Shah, N. H. (2013). Learning signals of adverse drug-drug interactions from the unstructured text of electronic health records. *AMIA Summits on Translational Science Proceedings AMIA Summit on Translational Science*, 83–87.
- Lependu, P., Iyer, S. V, Fairon, C., & Shah, N. H. (2012). Annotation Analysis for Testing Drug Safety Signals using Unstructured Clinical Notes. *Journal of Biomedical Semantics*, 3 Suppl 1(S5).
- Li, H. (2011). *Learning to Rank for Information Retrieval and Natural Language Processing. Synthesis Lectures on Human Language Technologies* (Vol. 4, pp. 1–113).
- Lin, W.-H., Wang, M.-C., Wang, W.-M., Yang, D.-C., Lam, C.-F., Roan, J.-N., & Li, C.-Y. (2014). Incidence of and mortality from Type I diabetes in Taiwan from 1999 through 2010: a nationwide cohort study. *PloS One*, 9(1), e86172.
- Lindquist, A. B. M., & Olsson, I. R. E. S. (1998). A Bayesian neural network method for adverse drug reaction signal generation. *Pharmacoepidemiology and Prescription*, 315–321.
- Liu, T.-Y. (2007). Learning to Rank for Information Retrieval. *Foundations and Trends® in Information Retrieval*, 3(3), 225–331.
- Park, M. Y., Yoon, D., Lee, K., Kang, S. Y., Park, I., Lee, S. H., ... Park, R. W. (2011). A novel algorithm for detection of adverse drug reaction signals using a hospital electronic medical record database. *Pharmacoepidemiology and Drug Safety*, 20, 598–607.
- Reps, J., Feyereisl, J., & Garibaldi, J. (2011). Investigating the detection of adverse drug events in a UK general practice electronic health-care database. *UKCI 2011, the 11th Annual Workshop on Computational Intelligence*, 167–173.
- Roughead, E. E. (1999). The nature and extent of drug-related hospitalisations in Australia. *Journal of Quality in Clinical Practice*, 19(1), 19–22.

- 
- Strom, B. L. (2012). Overview of Automated Databases in Pharmacoepidemiology. In *Pharmacoepidemiology* (pp. 158–162). Chichester, UK: John Wiley & Sons.
- Szarfman, A., Machado, S. G., & O’Neill, R. T. (2002). Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA’s spontaneous reports database. *Drug Safety : An International Journal of Medical Toxicology and Drug Experience*, 25(6), 381–392.
- Van der Hooft, C. S., Sturkenboom, M. C. J. M., van Grootheest, K., Kingma, H. J., & Stricker, B. H. C. (2006). Adverse drug reaction-related hospitalisations: a nationwide study in The Netherlands. *Drug Safety : An International Journal of Medical Toxicology and Drug Experience*, 29(2), 161–168.
- Wang, Y., He, D., & Chen, W. (2013). A Theoretical Analysis of NDCG Ranking Measures. *JMLR: Workshop and Conference Proceedings*, 1–30.
- Yu, H., & Kim, S. (2012). SVM Tutorial : Classification , Regression , and Ranking. In *Handbook of Natural Computing* (pp. 479–506). Springer Berlin Heidelberg.