

國立臺灣大學管理學院資訊管理學研究所



碩士論文

Department of Information Management

College of Management

National Taiwan University

Master Thesis

於健保資料庫中偵測藥物不良反應

Detecting Adverse Drug Reactions in

Health Insurance Claims Data

李承鑫

Cheng-Jen Lee

指導教授：盧信銘 博士

Advisor: Hsin-Min Lu, Ph.D.

中華民國 103 年 7 月

July 2014

致謝



此份論文得以圓滿完成，首先必須感謝我的指導教授盧信銘老師。盧老師總是耐心解釋令我眼花的數學模型，在我遭逢低潮信心盡失時，老師的鼓舞打氣讓我重獲動力；在我對未來徬徨無助時，老師中肯的分析讓我在做出抉擇時更加堅定。也要感謝李昇暉老師與曹承礎老師在口試時惠予的諸多寶貴建議，使我的論文在管理意涵上更臻完善。另外要特別感謝邱崇賢老師在學術英文寫作上的指點，以及莊庭瑞老師給予我研究助理的工作機會，使我得以進一步磨練程式能力。

我也要感謝實驗室的學長、同學及學弟妹們。我受到凱迪的許多鼓勵，和他交流對資訊產業的看法亦相當愉快；健華是我的數學小幫手，感謝他為我解開無數艱深難題；宇泰學長在生活和課程上的建議令我受益良多；與學弟妹們雖然相處時間不多但相談甚歡。同時我還要感謝親朋好友們。與易侁、玟郁、秉佳、誌軒學長及健彰學長的晚餐聚會，圍繞著各式御宅話題，好不快活；也感謝俊宇的協助，使論文得以順利付梓。另外還有來自虛擬世界的動畫、遊戲及社群網路上的朋友們，都是我在這段時間裡最佳的生活調劑。謝謝你們的鼎力相助。

最後，我要感謝我的家人。母親不時的電話關心，讓我重溫故鄉的美好；姑姑提供最好的資源使我得以專心於課業，更在生活上給予我許多幫助。謝謝你們的支持。「行百里，半九十」是我的人生座右銘。現在我又往前邁進一些，今後仍將繼續奮鬥，為更美好的未來打拼。ファイトだよ！

李承鑫 謹識

于臺大資訊管理學研究所

一百零三年七月

中文摘要



藥物不良反應 (Adverse Drug Reactions, 簡稱 ADRs) 係指接受藥物治療後所產生的嚴重健康危害。更由於 ADRs 是當今主要死因之一，故妥善監視上市後藥物成為一重要課題。然而，傳統的失衡分析法 (disproportionality analysis) 與貝氏偵測方法 (Bayesian signal detection) 仰賴預先收集的 ADR 通報案例，以及需事先定義、無統一標準的門檻值，其偵測結果也經常無法一致。另一方面，用以進行偵測的資料集長久受限於兩個資料庫—美國 FDA 之 FAERS 與 WHO 之 VigiBase，於這些資料庫的偵測也存在諸多困難。

為解決上述問題，本研究使用全民健康保險研究資料庫，以一週為單位聚合每位病患之歷史就診紀錄後，建立藥物與診斷先後關係。我們並提出一結合三種偵測分數：回歸 t 值 (*REG*)、通報相對比例值 (*PRR*) 與通報相對勝算比 (*ROR*) 作為輸入特徵的新模型，用以偵測藥物不良反應。實驗結果顯示，相較單獨使用一種分數，結合三種偵測分數的新模型之準確度 (Accuracy) 最高有 9.5% 的提升。

關鍵字：藥物不良反應、訊號偵測、健康資料庫、藥物安全監視、藥物主動監視

Abstract



Adverse Drug Reactions (ADRs) are fatal health problems due to medical treatments. ADRs are leading cause of death, and thus it is crucial to properly monitor post-marketing drugs. However, traditional disproportionality analysis and Bayesian signal detection depend on pre-collected ADR reports and a not universal, predefined threshold; the results are often inconsistent. Moreover, the available data sources were limited to two databases — U.S. FDA's FAERS and WHO's VigiBase; there are also several difficulties when detecting ADRs in these databases.

To address above problems, in this study, we proposed a model combining three detecting scores: regression's t-value (*REG*), proportional reporting ratio (*PRR*), and reporting odds ratio (*ROR*), as features for detecting serious drug-ADR pairs from one-week aggregated patient-week information with precedence relationship between drugs and diagnoses, in an health insurance claims database NHIRD (National Health Insurance Research Database). We demonstrated that the proposed combined score led to an improvement (up to 9.5%) of signal detection accuracy over applying each of score independently.

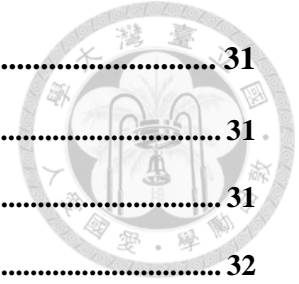
Keywords: adverse drug reaction, signal detection, administrative health database, drug safety surveillance, pharmacovigilance

Table of Contents



Table of Contents	i
List of Tables	iii
List of Figures	iv
Chapter 1 Introduction	1
Chapter 2 Literatures Review	4
2.1 ADR Signal Detection.....	4
2.1.1 Disproportionality Analysis (DPA)	6
2.1.2 Bayesian Signal Detection (BSD).....	7
2.1.3 The Problems of Traditional Approaches for ADR Detection	9
2.1.4 Other Approaches for ADR Detection	10
2.2 The Data Sources Used for ADR Detection.....	11
2.3 Evaluating Performance in ADR Detection	12
Chapter 3 Data and Models.....	15
3.1 Data Source	16
3.2 Patient Week Aggregation.....	18
3.3 Feature Generation.....	19
3.4 Reference Standard	24
3.5 Evaluation	25
Chapter 4 Results.....	27
4.1 Signaling Performance	27
4.2 Marginal Improvement of Combined Model.....	28
4.3 Differences between ADRs.....	29
4.4 The Effect of Period Length	29

Chapter 5 Conclusion.....	31
5.1 Contributions	31
5.2 Managerial Implication.....	31
5.3 Limitations and Future Work.....	32
References.....	33



List of Tables



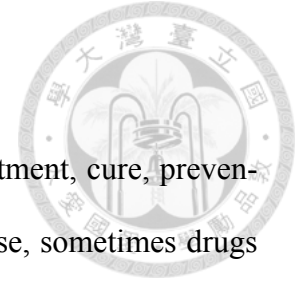
Table 1	Contingency Table for Detecting ADR.....	7
Table 2	Some DPA methods for detecting ADR.....	7
Table 3	Reference Standard, Performance Measures and Results in Previous Studies	13
Table 4	Data Tables and Fields Used in This Research.....	17
Table 5	Descriptive Statistics of The Dataset.....	18
Table 6	Reference Standard.....	24
Table 7	Experimental Settings.....	26
Table 8	Accuracies of Combined and Individual Scores by Three Classifiers.....	27
Table 9	F-measures of Combined and Individual Scores by Three Classifiers.....	28
Table 10	Marginal Improvement of Combined Model.....	29
Table 11	Average Performance of Six ADRs.....	29
Table 12	Accuracies for Combined Scores under Different Length of Period t	30

List of Figures



Figure 1	Processing pipeline for generating patient weeks, calculating scores and evaluating detected drug-ADR pairs	15
Figure 2	Illustration of aggregated diagnoses and drugs of two patients in the current period t and the past one period $t - 1$	20

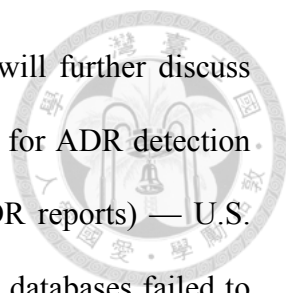
Chapter 1 Introduction



In pharmacology, drugs are crucial substances used in the treatment, cure, prevention, or diagnosis of disease. However, instead of treating the disease, sometimes drugs could harm patients or even threaten their life. Adverse Drug Reactions (ADRs, also called adverse events, or AEs) refer to fatal health problems, life-threatening events due to medical treatments (Harpaz, Vilar, et al., 2013; *Pharmaceutical data mining*, 2010).

U.S. FDA (Food and Drug Administration) reported that there were over 2 million serious ADRs yearly (Lazarou, Pomeranz, & Corey, 1998), at an annual cost of USD\$136 billion (Johnson, 1995). Moreover, a recent study estimated that in 2008 over 180,000 Americans would die from ADRs after using FDA-approved drugs, and ADRs are the sixth leading cause of death worldwide (Hacker, Messer, & Bachmann, 2009). To put this into perspective, consider that in 2008 there were around 120,000 Americans died from accidents, which was fewer than those caused by ADRs (Miniño, Murphy, Xu, & Kochanek, 2011). In Taiwan, it was reported that there were 831 ADR reports in 2013, an increase of 3500% over the last ten years (蔡雅婷, 陳文雯, & 蔡翠敏, 2014). However, former researches also showed that 42% life-threatening and serious ADRs were preventable with proper administration (Bates, 1995). Therefore, systematically tracking and validating ADRs are critical issues in both financial and social aspects.

In recent years, there has been increasing interest in using data mining techniques to automatically detect suspected ADRs. Organizations like WHO and U.S. FDA have built spontaneous reporting systems to record reported ADRs. They have also developed several statistical methods for routinely detecting possible ADRs from databases. However, we find that the methods introduced in former studies had the following problems when screening out possible ADRs: (1) easily affected by the number of reports, (2) do



not perform consistently, and (3) need predefined thresholds. We will further discuss them in Chapter 2.1.3. In addition, the choices of data sources used for ADR detection were limited to two reporting databases (database that collects ADR reports) — U.S. FDA’s FAERS and WHO’s Vigibase, and detecting from reporting databases failed to consider never reported drugs.

This study focuses the following research questions:

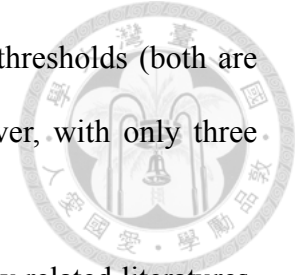
1. Is it possible to develop a new model with good performance and address the existed issues in ADR detection?
2. Is there any other source suitable for ADR detection? How should these data be adjusted for detection?

Therefore, in this research, we proposed a new methodology, which combined three scores (regression’s t-value, proportional reporting ratio, and reporting odds ratio) from aggregated patient-week information and put the scores into three classifiers (RBF-SVM, random forest, and logistic regression) for detecting possible drug-ADR pairs in the Taiwan’s National Health Insurance claims database (NHIRD).

Six known ADR groups in three categories: Cardiovascular Disease, Hepatotoxicity, and Cancer were selected for investigation. The evaluation process is expressed as a classification problem with three classes, including one serious type and other two types of drug-ADR pairs. We argued that our model with combined scores outperforms individual scores in separating serious type of pairs from other pairs when evaluating in accuracy, precision, recall, F-measure.

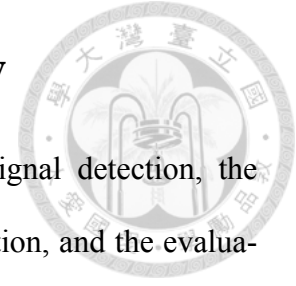
This research differs from the previous research that ran detection processes using traditional disproportionality analysis or Bayesian signal detection on reports-based databases (A Bate & Evans, 2009; Dumouchel, 1999; Kubota, Koide, & Hirai, 2004; *Pharmaceutical data mining*, 2010). Since we detect ADRs through a supervised classi-

fication on a health database, neither ADR reports nor predefined thresholds (both are required in traditional methods) are needed in our model. Moreover, with only three simple scores, our model is easy to implement and time-efficient.



This paper is organized as follows. In chapter 2, we will review related literatures. Chapter 3 will describe the data, methodology, and evaluation method used in this paper. Chapter 4 will report and discuss the results. Finally, we will make a short conclusion in Chapter 5.

Chapter 2 Literatures Review



Our study covers several research issues, including ADR signal detection, the methods used for signal detection, the data sources for signal detection, and the evaluation processes. In this chapter, we will discuss the related literatures for them.

2.1 ADR Signal Detection

Signal detection, or drug safety, is defined as a series of activities for understanding and preventing adverse drug reactions (World Health Organization & WHO Collaborating Centre for International Drug Monitoring, 2002). To record and trace ADRs, WHO requires drug firms and medical staff to report possible or confirmed ADRs for post-marketing drugs, and then gathered the drug name, used record, diagnosis and basic information about the patient into the VigiBase, an electronic reporting database developed and maintained by WHO UMC (Uppsala Monitoring Centre) (Norén, Sundberg, Bate, & Edwards, 2008). U.S. FDA also started a similar system in 1969, and stored the collected data in their FAERS (FDA Adverse Event Reporting System) database (Dumouchel, 1999). As of the third quarter of 2010, there had been more than 4 million reports on FAERS (Harpaz, Vilar, et al., 2013); and in April 2013, the number of cases in the WHO VigiBase reached 8 million (Uppsala Monitoring Centre, 2013).

Through the adoption of reporting databases, it is expected that ADRs can be detected as early as possible (Cornelius, Sauzet, & Evans, 2012; Dumouchel, 1999), and then the firms can improve their products, or the authorities can withdraw the problematic drugs (*Pharmaceutical data mining*, 2010). This may minimize the impact of ADRs on patients (Harpaz, Chase, & Friedman, 2010; Jha et al., 1998).

The most important task on the reporting databases is to detect suspicious ADR

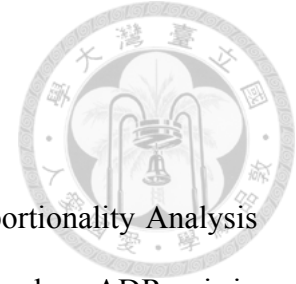
“signal” on reporting data through its properties such as its severity and reporting numbers (Harpaz, Vilar, et al., 2013; *Pharmaceutical data mining*, 2010). WHO defines the process of signal detection as “Reported information on a possible causal relationship between an adverse event and a drug, the relationship being unknown or incompletely documented previously.” Though the above vision is attractive, there are several limitations when conducting ADR signal detection using reporting databases (A Bate & Evans, 2009; *Pharmaceutical data mining*, 2010).

Firstly, to obtain robust and comprehensive result, experiments on large datasets for the long period of time are required. Bright *et al.* assessed five CDSS (Clinical Decision Support System) studies and showed that neither these studies had significant effect on reducing ADRs due to deficient evaluation period (Bright et al., 2012). However, it is nearly impossible to manually detect all possible ADR signals with limited time and human power. To detect signals more effectively, a series of quantitative approaches were constructed (A Bate & Evans, 2009). The main idea behind these approaches is to monitor and detect drug-ADR or drug-symptom patterns. These approaches will be introduced in the following paragraphs.

Secondly, the detecting process may filter out clinically insignificant ADR alerts that domain experts may ignore (Kuperman et al., 2007), which may result in useless detection results. Hence, it is crucial to evaluate the system performance carefully.

Finally, the detecting process can work only when patients’ health records, such as diagnosis and drug records, are fully entered into the databases (Kuperman et al., 2007).

On the other hand, the other information provided by databases is also valuable. They are: patients’ demographics, longitudinal data for drug usage, and institutions’ information, etc. For example, we can perform stratum specific estimates, and thus prevent from misleading with a certain stratum (A Bate & Evans, 2009).



2.1.1 Disproportionality Analysis (DPA)

The most widely used procedure of detecting ADR is Disproportionality Analysis (DPA). The concept behind DPA is accessing the “relative risk” of a drug-ADR pair in comparison with other pairs. However, due to the lack of accurate denominator (i.e., number of doses of administered drug), it is hard to identify the potential combinations from the spontaneous reports. Thus, in DPA method, there is an assumption of “baseline frequencies” (or baseline risk) as denominator.

We now focus on the 2x2 contingency table for drugs and ADRs, as shown in Table 1. Take relative report rates (RR) for example. The expected baseline frequencies are $[(a + c) \times (a + b)] / (a + b + c + d)$ when reports involving focused drug are statistically independent of reports involving focused ADR. Intuitively, we can decide whether focused drug-ADR pair is suspected by evaluating the ratio of its number of reports (a) and the corresponding baseline frequencies $[(a + c) \times (a + b)] / (a + b + c + d)$. Hence, the RR is $[a \times (a + b + c + d)] / [(a + c) \times (a + b)]$. In practice, we first calculate the degree of disproportionality through a predefined formula (e.g., RR), and then examine the confidence interval (CI) for disproportionality. If the lower limit of CI is higher than a given threshold, the relationship between the drug and the ADR is considered suspected.

Table 2 demonstrates the formula for calculating disproportionality and threshold under each DPA method, including RR, PRR (Proportional Reporting Ratio), ROR (Reporting Odds Ratio), and MCA (a comprehensive metric suggested by UK’s Medicines and Healthcare Products Regulatory Agency) (Kubota et al., 2004).

Table 1 Contingency Table for Detecting ADR

	Reports of focused ADR	Reports of other ADRs	Total
Reports of focused drug	a	b	a + b
Reports of other drugs	c	d	c + d
Total	a + c	b + d	a + b + c + d

Table 2 Some DPA methods for detecting ADR

Method	Formula	Threshold	Adoption
RR	$\frac{a(a + b + c + d)}{(a + c)(a + b)}$	No unified threshold.	
PRR	$\frac{a/(a + b)}{c/(c + d)}$	$PRR - 1.96SE > 1$	Netherlands PFL*
ROR	$\frac{a/c}{b/d}$	$ROR - 1.96SE > 1$	
MCA	PRR, a, χ^2	$PRR \geq 2, a \geq 3, \text{ and } \chi^2 \geq 4$	UK's MHRA**

* Pharmacovigilance Foundation Lareb

** Medicines and Healthcare Products Regulatory Agency

2.1.2 Bayesian Signal Detection (BSD)

Another popular choice for detecting ADRs is Bayesian Signal Detection (BSD). BSD is actually an extension of DPA with Bayesian inference.

Although DPA is a simple method, its results are easily affected by the number of reports in Table 1. Take RR for example. When total amount of drug-ADR combinations $N = a + b + c + d$ is very large (e.g., $N = 1,000,000$) but the amount of reports is relatively small (e.g., $a = 1$), the baseline frequencies of reports in RR may be considerably small. Small baseline frequencies would lead to fairly large RR value. However, a combination with relatively small count means lower support, thus its RR value is meaningless. Under traditional DPA framework, these meaningless drug-ADR combina-

tions will be wrongly screened out (Dumouchel, 1999).

U.S. FDA has deployed Empirical Bayes Gamma-Poisson Shrinker (EBGPS) for routine ADRs screening to find whether a combination of drug and ADR is frequently reported than expected. By introducing gamma prior, EBGPS method shrinks RR and then addresses above problem. EBGPS utilizes Bayes' theorem to obtain the ratio of observation to expected number of reports for a combination. For a focused drug-ADR combination (drug i , ADR j), the statistic is defined below:

$$\text{EBGPS}_{ij} = 2^{E[\log(\lambda)|N=N_{ij}]/\log(2)}$$

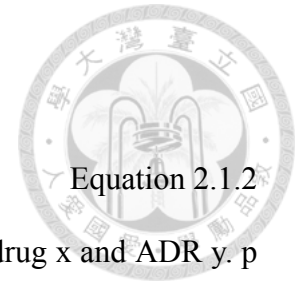
Equation 2.1.1

$\lambda = \frac{\mu_{ij}}{E_{ij}}$ denotes the observed count of reports for (i, j) drawn from a Poisson distribution with a mean μ_{ij} . In the λ , $E_{ij} = \sum_k N_{i,k} N_{j,k} / N_{..k}$ denotes the expected amount of reports for a drug-ADR combination (i, j) , where $N_{i,k} = \sum_j N_{ijk}$ and $N_{j,k} = \sum_i N_{ijk}$ denote the total number of observed reports involving drug i and j under a stratum (includes time and age) k respectively, whereas $N_{..k} = \sum_i \sum_j N_{ijk}$ denotes total number of observed reports belonging to the stratum k .

If the lower limit of the 95% CI of EBGPS_{ij} is higher than a threshold (0.5 as recommended by the FDA) (Deshpande, Gogolak, & Smith, 2010), then the combination (i, j) is considered a possible drug-ADR pair (A Bate & Evans, 2009; Dumouchel, 1999).

WHO screens out ADRs through Bayesian Confidence Propagation Neural Network (BCPNN) (Andrew Bate, Lindquist, Edwards, & Orre, 2002; Orre, Lansner, Bate, & Lindquist, 2000), which is done by evaluating information component (IC) (in fact, IC itself is another DPA measure):

$$IC = \log_2 \frac{p(x, y)}{p(x)p(y)} = \log_2 \frac{a(a + b + c + d)}{(a + b)(a + c)}$$



Equation 2.1.2

IC can be used to examine the degree of relationship between drug x and ADR y . $p(x)$ and $p(y)$ stands for the proportion of drug x and ADR y in all reports, respectively. The higher IC means the stronger relationship between drug x and ADR y . If the lower limit of the 95% confidence interval of the Bayesian posterior (through Bayesian inference. we omitted the inference here.) of IC is higher than 0 (Deshpande et al., 2010), then the combination is considered a possible drug-ADR pair.

2.1.3 The Problems of Traditional Approaches for ADR Detection

Both DPA and BSD are widely used methods for ADR detection. However, there are several problems in the two traditional approaches.

First, for DPA, we have known that the effect of small number of reports may generate bad drug-ADR combinations through the discussion in the beginning of 2.1.2. Second, every traditional method requires a predefined threshold, which is decided by specific data source. For DPA, the last column in Table 2 shows the data source; for BSD, the thresholds of EBGPS and BCPNN are decided according to the data stored in FAERS and VigiBase, respectively (Deshpande et al., 2010). However, a threshold designed for one data source may be unsuitable for other sources. Moreover, when N in Table 1 equals 1 or 2, former studies found that the Kappa statistics between DPA, BCPNN and EBGPS are small, showing the inconsistency among traditional methods (Kubota et al., 2004).

Therefore, an ideal approach for ADR detection should be a simple and well-performed model without the need for reports and thresholds. In addition, building

a model that combines multiple detection methods may reduce the inconsistency among methods and then improve the performance.



2.1.4 Other Approaches for ADR Detection

Other than traditional DPA and Bayesian methods, there are still many approaches for detecting ADRs in electrical databases. They may be an extension of DPA or Bayesian concept, or based on other data mining techniques, such as frequent pattern mining.

Norén *et al.* proposed a disproportionality measure, which is based on a baseline model with additive risk, for exploratory analysis of suspected drug-drug interaction (DI) in VigiBase (Norén et al., 2008). They provided examples and argued that this modified DPA method can detect more DIs in comparison with the model using third-order log-odds ratio. Choi *et al.* applied this measure to the Korean National Health Insurance claims database and successfully screened out an actual DI between two drugs (Choi, Chang, Choi, Chung, & Shin, 2013). However, both of the above two studies failed to show the credibility of this measure due to limited numbers of case studies.

Jin *et al.* utilized the concept of temporal association rule and developed a mining algorithm MUTARA (Mining Unexpected Temporal Association Rules given the Antecedent) to highlight the “unexpected drug-to-diagnosis patterns” (Jin et al., 2010). This algorithm finds the unexpected patterns by removing the “expected diagnoses” occurred before focused drug, in the time-constrained hazard period. They also proposed a measure called rankRatio, which combined the rank under traditional temporal association rule and that under their unexpected temporal association rule. They showed that MUTARA and rankRatio could signal more ADRs than traditional temporal association rule mining techniques.

2.2 The Data Sources Used for ADR Detection

Although there is no shortage of methods to detect possible ADRs, the choices of databases used to analyze are limited. U.S. FDA's FAERS and WHO's Vigibase, the two databases mentioned above, are the main data sources adopted in previous studies (Harpaz et al., 2010). Also, this kind of reporting system has several shortcomings leading to difficulties in analysis. Firstly, lack of the number of drug uses at a specific time and patient makes it impossible to estimate incidence density and risk. Secondly, clinicians may underreport the frequency of ADRs due to extra workload, insufficient attention for ADR, and fear of lawsuits (Jin et al., 2010). Finally, these systems require direct reports from medical staff, thus delaying the detecting process (Cornelius et al., 2012).

On the other hand, recently, population-based administrative health databases like health insurance claims databases and electronic health records (EHR) have become popular choices for ADR signal detection (Coloma et al., 2012; Johansson, Wallander, de Abajo, & García Rodríguez, 2010; Park et al., 2011). Compared to traditional reporting databases, population-based databases have more comprehensive information for all patients, whether they were exposed to the drug or not (Cornelius et al., 2012; Sauzet, Carvajal, Escudero, Molokhia, & Cornelius, 2013). This makes it possible to detect signals for those drugs not included in the traditional reporting databases, and to make longitudinal studies (Cornelius et al., 2012).

Jha *et al.* established a rule-based monitoring program, which contained 52 unique rules (Most of the rules are laboratory abnormalities), to find ADRs from a hospital's clinical-results reporting system. They showed that this computer monitoring strategy not only had acceptable capture rate (45% vs. 65%) on ADRs but required fewer people in-

involved (11 person-hours vs. 55 person-hours) when compared with traditional chart review (Jha et al., 1998).

Sauzet *et al.* applied the Weibull Shape Parameter (WSP) test (a time-to-event model for finding the time to high risk ADRs) (Cornelius et al., 2012) to the Health Improvement Network (THIN), an EHR database in UK, and successfully detected two well-known ADRs (Sauzet et al., 2013).

Harpaz *et al.* tried to combine the detected signals using U.S FAERS and the one using narratives processed with natural language processing (NLP). They found that the combined signals showed better precision, including precision at K and F-measure (Harpaz, Vilar, et al., 2013).

Overall, administrative health database is a potential resource for ADR detection due to its better availability and richer content. Hence, in this study, we want to detect ADR signals over a long period of time in a large administrative health database.

2.3 Evaluating Performance in ADR Detection

Evaluating the performance is another important and challenging issue in ADR detection. Table 3 lists the reference standard, measures used in former studies, accompanied with primary results.

Table 3 Reference Standard, Performance Measures and Results in Previous Studies

#	Study	Performance Measures and Results	Reference Standard or Evaluation Process
1	Jha et al. (1998)	PPV*: 0.17	Domain experts reviewed the results.
2	Dumouchel (1999)		Compare the ranking.
3	Norén et al. (2008)		Compare the ranking.
4	A Bate & Evans (2009)	PPV: 0.44 NPV: 0.85	Retrospective evaluation by using ADRs listed in 2000
5	Jin et al. (2010)	Accuracy: 0.313	Domain experts reviewed the results.
6	Cornelius et al. (2012)	Accuracy: 0.53~0.93 Sensitivity: 0.21~1.0 Specificity: 0.81~0.87 False positive rate: 0.07~0.12 False negative rate: 0~0.39	30,000 simulated datasets for ADRs across time
7	Harpaz et al. (2013)	AUC***: 0.76~0.94	380 positive and negative cases (drug-outcome)
8	Harpaz, Vilar, et al. (2013)	Precision at K: 0.27~0.85 Recall at K: 0.15~0.2 F-measure: 0.17~0.31	Known drug-ADR pairs

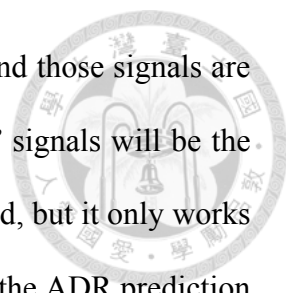
* Positive predictive rate

** Negative predictive rate

*** Area under the curve of ROC (receiver operating characteristic)

Results are in a range for different parameter settings.

Overall, there are three types of reference standard used in ADR detection. The first is a prepared list of known ADRs, which had been used in study #6, #7 and #8. This kind of reference is easy to build, but the results of evaluation will be limited to the reference list. The second is a retrospective evaluation, which defines an ADR as a “signal” when it is detected from historical data and newly appears at a later time. For



instance, study #4 detected ADR signals from the reports in 1993, and those signals are compared with all signals newly appeared in 2000. Then, the “true” signals will be the matched ones. Retrospective evaluation is reasonable and widely used, but it only works on a dataset with temporal information. The third is directly sending the ADR prediction to domain experts for further examination. Study #1 and #5 are two examples. This method is the most straightforward and reliable; however, this evaluation process involves human power, thus increasing the screening time.

The majority of performance measures involve metrics of a classification test, including accuracy, precision (positive predictive rate, or the PPV), recall (sensitivity), specificity (true negative rate), F-measure, false positive rate, AUC (area under the curve of receiver operating characteristic), etc. From the third column in Table 3, we have a general idea about how the former screening performed despite the fact that the results vary under different settings, dataset, and evaluation process.

Note that instead of using above metrics, studies #2 and #3 compared the top-ranked signals detected by proposed method and baselines. This method provides another choice for evaluation when there exist significant difference in ranking results between methods.

Chapter 3 Data and Models



This chapter will discuss the dataset and methodology used in this research. Our work is motivated by the belief that there exists precedence relationship for a patient among history of drug use and the diagnoses happened afterwards. A series of patient visits from Taiwan's National Health Insurance Research Database (NHIRD) will be aggregated to obtain drugs and diagnoses information of each patient in each one-week period (in Chapter 3.2). Then these patient weeks will be used for calculating regression's t-value, PRR, and ROR score for each drug-diagnosis pair (in Chapter 3.3). Finally, these three scores will be combined and provided as features to detect serious drug-ADR pairs using three classification algorithms (in Chapter 3.5). Figure 1 provides an outline of the process.

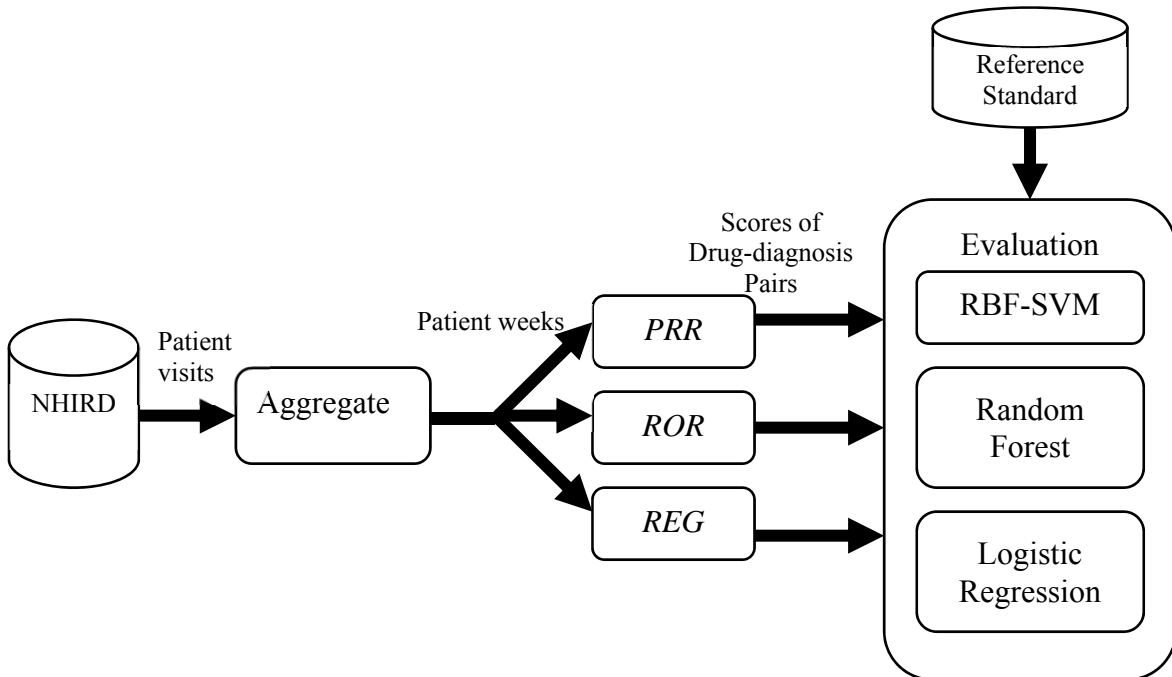


Figure 1 Processing pipeline for generating patient weeks, calculating scores and evaluating detected drug-ADR pairs

3.1 Data Source

Our study adopted NHIRD, which is an administrative health database contains healthcare service claims submitted to National Health Insurance Administration (NHI) by healthcare providers. The data include information on medical treatments and diagnoses for both inpatients and outpatients.

We extracted three years of inpatient and outpatient data (from 23 December 2007 to 1 January 2011) from the NHIRD. We only considered patients older than age 20. Data tables and fields used in this research are shown as Table 4. Note that due to the lack of definition of visit time in the inpatient data (DD table), we used date of admission (IN DATE) here. Similarly, prescribing date (FUNC DATE) is used to represent date of visit in the contracted pharmacies data (GD table). We included contracted pharmacies here because in many medical facilities (e.g., local clinics) drugs are prescribed by medical staff and then made up by contracted pharmacies.

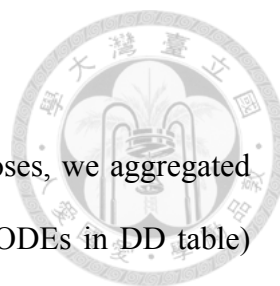
Diagnoses were encoded according to the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) system. For drugs, since they were encoded according to NHI's self-defined coding schemes, we mapped them to the WHO Anatomical Therapeutic Chemical (ATC) system. The mapping table was retrieved from the NHI's website (National Health Insurance Administration, Ministry of Health and Welfare, 2014) and examined by domain experts.



Table 4 Data Tables and Fields Used in This Research

Data Table	Field	Meaning
2000 Registry for Beneficiaries (ID2000)	ID_BIRTHDAY ID_SEX	For controlling confounding factors.
Ambulatory Care Expenditures by Visits (CD)	FUNC_DATE	Date of visit
	ACODE_ICD9_1 ACODE_ICD9_2 ACODE_ICD9_3	Diagnosis
Details of Ambulatory Care Orders (OO)	DRUG_NO	Drug code*
Inpatient Expenditures by Admissions (DD)	IN_DATE	Date of visit
	ICD9CM_CODE ICD9CM_CODE_1 ICD9CM_CODE_2 ICD9CM_CODE_3 ICD9CM_CODE_4	Diagnosis
Details of Inpatient Orders (DO)	ORDER_CODE	Drug code*
Expenditures for Prescriptions Dispensed at Contracted Pharmacies (GD)	FUNC_DATE	Date of visit
Details of Prescriptions Dispensed at Contracted Pharmacies (GO)	DRUG_NO	Drug code*
2000 Registry for drug prescriptions (DRUG2000)	DRUG_ID	Definition of drug codes

* Drug codes have been mapped to ATC.



3.2 Patient Week Aggregation

To build the precedence relationships among drugs and diagnoses, we aggregated the diagnoses (from `ACODE_ICD9s` in `CD` table and `ICD9CM_CODES` in `DD` table) and prescribed drugs (from `DRUG_NO` in `OO` table, `ORDER_CODE` in `DO` table, and `DRUG_NO` in `GO` table) in every patient visit according one-week time period. We selected seven days as time period because we found that the choice of period length has no significant effect on performance (we will discuss more in Chapter 4.4.) and aggregation by shorter period may generate more patient visits (i.e., more training examples). The term *patient week* is used to depict a set of aggregated visits of a patient in a one-week period. Moreover, we removed drugs prescribed less than 30 times in all periods.

With the above processing, as shown in Table 5, we got approximately 7.5 million aggregated patient weeks accompanied with diagnoses, prescribed drugs, and basic patient information (age and sex) for 614,080 patients in 158 time periods. We will use these aggregated patient weeks as our dataset in our study. Table 5 also shows other descriptive statistics of our dataset. The preprocessing was completed using a PostgreSQL 9.1 database system (available at www.postgresql.org).

Table 5 Descriptive Statistics of The Dataset

Number of unique diagnoses	1,125
Number of unique drugs (encoded by ATC)	1,326
Number of patient weeks	7,534,707
Number of unique patients (age \geq 20)	Male: 326,587 Female: 287,493 Total: 614,080

Mean/Median age of patients	54.1/54
Max/Min age of patients	107/20
Mean of diagnoses per patient week	2.24
Standard deviation of diagnoses per patient week	1.48
Mean of prescribed drugs per patient week	4.24
Standard deviation of prescribed drugs per patient week	3.39

3.3 Feature Generation

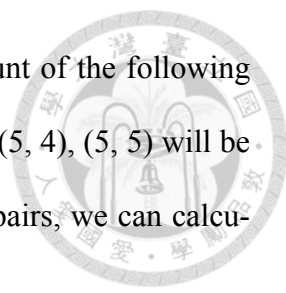
The following notation is useful:

- $t = 1, 2, \dots T$: The time period index.
- $p = 1, 2, \dots P$: The patient index.
- $m = 1, 2, \dots M$: The focused diagnosis index.
- $n = 1, 2, \dots N$: The focused drug index.

Therefore, a focused drug-diagnosis pair can be expressed as (n, m) .

Now we can calculate three “scores” for each drug-diagnosis pair: (1) *PRR* by the formula of *PRR*, (2) *ROR* by the formula of *ROR*, and (3) regression’s t-value. We select these three scores for two reasons. Firstly, *PRR* and *ROR* have been widely used to find the relationship between drug and diagnosis, and the regression’s t-value has been a popular metric for measuring relationships among variables as well. Secondly, all the three scores are easy to calculate through a predefined formula (for *PRR* and *ROR*) or a simple closed-form solution to the model (for regression).

For the first two scores: *PRR* and *ROR*, the count of a focused (n, m) was generated as cumulative number of diagnosis m happened in the current period t after using drug n in the past one period $t - 1$ for all patients p in all periods t . Take patient 1 in Figure 2 for example, there are three diagnoses $m = 3, 4, 5$ in current period t and



three drugs $n = 2, 3, 5$ in past one period $t - 1$; therefore, the count of the following drug-diagnosis pairs: $(2, 3), (2, 4), (2, 5), (3, 3), (3, 4), (3, 5), (5, 3), (5, 4), (5, 5)$ will be increased by one. After accumulating the counts of drug-diagnosis pairs, we can calculate *PRR* and *ROR* using the formula listed in Table 2.

To illustrate, if there are only four patient weeks as shown in Figure 2, for a focused drug-diagnosis pair $(5, 3)$, The contingency table (Table 1) can be filled with $(a = 2, b = 2, c = 3, d = 4)$. Hence, the *PRR* for pair $(5, 3)$ will be $[2/(2+2)]/[3/(3+4)] = 1.17$ and the *ROR* for pair $(5, 3)$ will be $[2/3]/[2/4] = 1.33$.

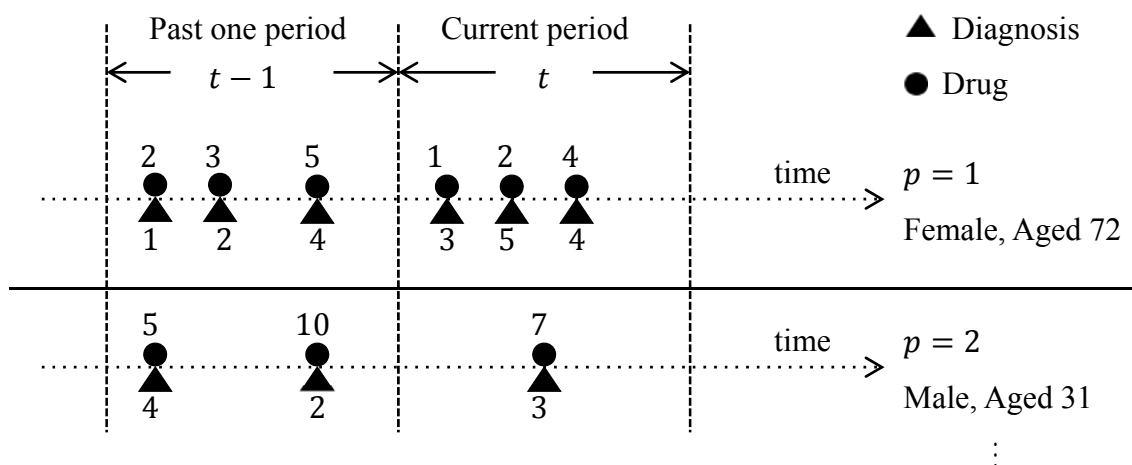


Figure 2 Illustration of aggregated diagnoses and drugs of two patients in the current period t and the past one period $t - 1$.

For the regression's t-value score, we will conduct a multivariate linear regression to obtain the t-value of each drug-diagnosis pair. We first let

- $\mathbf{R}_{t,p} = (r_1, r_2, \dots, r_m, \dots, r_M)$: A vector of M elements. Each r_m indicates whether a patient p has the correspond **diagnosis** m during period t .
- $\mathbf{D}_{t,p} = (d_1, d_2, \dots, d_n, \dots, d_N)$: A vector of N elements. Each d_n indicates whether a patient p uses the correspond **drug** n during period t .

We assume that for a patient p , the diagnoses happened during current period t (i.e., $\mathbf{R}_{t,p}$) is determined by drugs used in the past one period $t - 1$ (i.e., $\mathbf{D}_{t-1,p}$). However, consider that current drugs $\mathbf{D}_{t,p}$ may affect current diagnoses $\mathbf{R}_{t,p}$, to control this effect, we also added $\mathbf{D}_{t,p}$ to our regression model; $\mathbf{R}_{t-1,p}$ was added to the model for the same reason. Finally, to control confounding factors, all patients were stratified based on their age and gender, and were denoted by

$$A_{t,p} = \begin{cases} 0, & \text{age}(p \text{ during } t) < 60 \\ 1, & \text{age}(p \text{ during } t) \geq 60 \end{cases} \text{ and } S_p = \begin{cases} 0, & p \text{ is a female} \\ 1, & p \text{ is a male} \end{cases} \text{ respectively.}$$

In the research dataset, there are 4,562,381 patient weeks whose patients aged under 60 (but over 20 by settings) and 2,972,326 patient weeks whose patients aged over 60.

Then, the precedence relationship between $\mathbf{R}_{t,p}$ and $\mathbf{D}_{t-1,p}$ can be represented as following multivariate regression:

$$\mathbf{R}_{t,p} = \boldsymbol{\beta}'_a \mathbf{D}_{t-1,p} + \boldsymbol{\beta}'_b \mathbf{R}_{t-1,p} + \boldsymbol{\beta}'_c \mathbf{D}_{t,p} + \beta_d A_{t,p} + \beta_e S_p + \varepsilon_{t,p}$$

Equation 3.3.1

where $a = [1, N]$, $b = [N + 1, N + M]$, $c = [N + M + 1, 2N + M]$, $d = 2N + M + 1$, and $e = 2N + M + 2$ denote the indexes for model parameter $\boldsymbol{\beta}$. $\varepsilon_{t,p}$ is a zero mean Gaussian random variable with variance σ^2 .

To illustrate, for Patient $p = 1$ in Figure 2, a drug-diagnosis relationship can be represented as $(\mathbf{R}_{t,1}, \mathbf{D}_{t-1,1}, \mathbf{R}_{t-1,1}, \mathbf{D}_{t,1}, A_{t,1} = 1, S_1 = 0)$, where

$$\mathbf{R}_{t,1} = (r_3 = r_4 = r_5 = 1, \text{others} = 0),$$

$$\mathbf{D}_{t-1,1} = (d_2 = d_3 = d_5 = 1, \text{others} = 0),$$

$$\mathbf{R}_{t-1,1} = (r_1 = r_2 = r_4 = 1, \text{others} = 0), \text{ and}$$

$$\mathbf{D}_{t,1} = (d_1 = d_2 = d_4 = 1, \text{others} = 0).$$

To simplify the expression, we put the independent variables $(\mathbf{D}_{t-1,p}, \mathbf{R}_{t-1,p}, \mathbf{D}_{t,p}, A_{t,p}, \text{ and } S_p)$ together as one vector \mathbf{X} , which has $2N + M + 2$ variables. Then Equation 3.3.1 can be written in the following form:

$$\mathbf{R} = \boldsymbol{\beta}'\mathbf{X} + \boldsymbol{\varepsilon}$$

Equation 3.3.2

Through Maximum Likelihood Estimation, we obtain the optimal $(2N + M + 2)$ -dimensional parameter vector $\boldsymbol{\beta}_m$ for each $\mathbf{r}_m \in \mathbf{R}$,

$$\boldsymbol{\beta}_m = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{r}_m$$

Equation 3.3.3

And, for the model stability, we added a tiny smoothing constant 0.001 in every element on the main diagonal of $\mathbf{X}'\mathbf{X}$.

For convenience, we define S_{qq} as the q -th element (independent variable) on the main diagonal of $(\mathbf{X}'\mathbf{X})^{-1}$. Then the variance of $\boldsymbol{\beta}_m$ can be expressed as:



$$\begin{aligned}
\text{Var}(\boldsymbol{\beta}_m) &= (\mathbf{X}'\mathbf{X})^{-1}\sigma_m^2 \\
&= \sigma_m^2 \begin{bmatrix} S_{11} & \cdots & \cdots & \cdots & \cdots \\ \vdots & S_{22} & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & S_{qq} & \vdots \\ \cdots & \cdots & \cdots & \cdots & S_{(2N+M+2)(2N+M+2)} \end{bmatrix} \\
&= \begin{bmatrix} \sigma_m^2 S_{11} & \cdots & \cdots & \cdots & \cdots \\ \vdots & \sigma_m^2 S_{22} & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \sigma_m^2 S_{qq} & \vdots \\ \cdots & \cdots & \cdots & \cdots & \sigma_m^2 S_{(2N+M+2)(2N+M+2)} \end{bmatrix}
\end{aligned}$$



Equation 3.3.4

where $\sigma_m^2 = \frac{\boldsymbol{\varepsilon}_m' \boldsymbol{\varepsilon}_m}{TP - (2N+M+2)}$ is the variance of error variable $\boldsymbol{\varepsilon}_m = \mathbf{r}_m - \mathbf{X}\boldsymbol{\beta}_m$, $\boldsymbol{\varepsilon}_m \sim \mathcal{N}(0, \sigma_m^2 \mathbf{I})$.

Therefore, to realize how does a diagnosis m be influenced by a drug n , we can calculate t-value $t_{m,n}$ by:

$$\frac{\beta_{m,n}}{\sqrt{\sigma_m^2 S_{nn}}}$$

Equation 3.3.5

If a certain drug n has higher $t_{m,n}$, it may have more influence on diagnosis m . And the drug-diagnosis pair (n, m) may be considered as a possible ADR signal. We hereinafter used the term *REG* to represent the calculated t-values through regression.

All the three scores (*REG*, *PRR*, and *ROR*) were calculated using statistical software R (available at www.r-project.org). Given computed *REG*, *PRR*, and *ROR* to all drug-diagnosis pairs, our methodology is to combine all the three scores as features and put them into a supervised classification algorithm to detect serious drug-ADR pairs from those candidate pairs.



3.4 Reference Standard

To evaluate the performance, we constructed a reference standard by labeling known drug-ADR pairs for six ADR groups in three categories: Cardiovascular Disease (CV), Hepatotoxicity (HEP), and Cancer (CAN) with drugs that may induce the ADRs. We combined these ADRs and drugs and then used these drug-ADR pairs for the reference standard in this study. Our reference standard dataset covers 75 ICD-9-CM codes and 553 drugs.

The domain experts classified a drug-ADR pair into six types: Type I (indications; valid usages to use drug for diagnosis), A (validated by large-scale experiments), B (validated by case reports), C (validated by animal studies), D (carried in leaflets of drugs), and no-relations. The reference standard is shown as Table 6 (we omitted the information of no-relation pairs).

Table 6 Reference Standard

ADR Cat.	ADR Group	ICD-9-CM	Num. of Drugs	Number of Drug-ADR Pairs				
				I	A	B	C	D
CV	Myocardial Infarction (MI)	410, 411	56	44	56	6	0	12
	Angina (ANG)	413, 414	31	28	26	12	0	2
	Arrhythmia (ATA)	426, 427	139	24	154	38	0	37
	Congestive Heart Failure (CHF)	402, 404, 428	50	18	96	15	0	24
HEP	Hepatotoxicity (HEP)	570, 573, 576	244	60	63	396	0	222
CAN	Cancer (CAN)	140 to 208 (63 in total)	33	1575	504	0	63	0
Total			553	1,749	899	467	63	297

Note: The combinations of each drug-ADR pair do not equal the sum of pairs for each type, because some pairs were identified by more than one type.

3.5 Evaluation

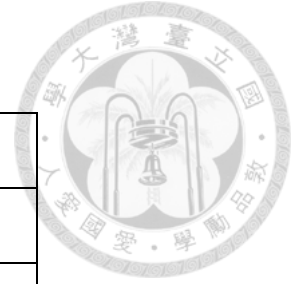
Through the discussion in Chapter 3.4, apparently, only Type A, B, and C are the reasonable “serious” drug-ADR pairs we should focus on. Hence, the evaluation process can be expressed as a classification problem with three classes, including one serious type of pairs (consist of Type A, B and C) and other two types of pairs (Type I and D respectively). As shown in Table 6, there are 1,429 (899 + 467 + 63) serious pairs, 1,749 type I pairs, and 297 type D pairs on the reference list.

With a prepared reference list, our evaluation process is simple and efficient in comparison with retrospective evaluation and expert examination. Additionally, because we only consider some serious drug-ADR pairs, it will not be a problem to limit detection results within our reference list.

We compared the proposed combined model with models that use *PRR*, *ROR*, and *REG* for feature independently on classification performance. The experimental settings are shown as Table 7. Among four feature settings, #4 is the proposed setting and #1~#3 are control groups with single feature. We put the four feature settings into three classification algorithms: random forest, RBF-kernel SVM, and logistic regression, respectively. By testing multiple classifiers, the bias of specific classifier should be prevented. R package randomForest, e1071, nnet are used for random forest, SVM, and logistic regression modeling respectively. In addition, for all algorithms, the evaluation was done by 10-fold cross-validation based on all test cases; for RBF-kernel SVM, a 10-fold cross-validation was conducted to select the best model parameters C (cost) and γ in the radial basis function.

Table 7 Experimental Settings

#	Feature Setting
1	<i>REG</i>
2	<i>PRR</i>
3	<i>ROR</i>
4	Combined (<i>PRR + ROR + REG</i>)



Performance was measured using accuracy, macro-average precision, macro-average recall, and macro-average F-measure. We considered the model with higher classification performance a better model for detecting serious drug-ADR pairs. Note that here we used the macro-average measures rather than ordinary ones because there were three classes to predict in this classification problem; moreover, we omitted the micro-average ones because in this problem each sample is always classified to belong to one of classes and therefore micro-averages are exactly the same as macro-averages.

Chapter 4 Results



4.1 Signaling Performance

Firstly, we report that the proposed methodology can reliably detect potential serious ADR signals among the drug-diagnosis pairs with different classifiers. Table 8 demonstrates comparison of the signal detection (classification) accuracy across individual scores and combined score by using three classifiers. Overall, the accuracies of combined scores are better than that of individual scores when evaluating by RBF-SVM and random forest; when it comes to logistic regression, the accuracies of combined scores are nearly the same as that of *PRR* and *ROR*, and better than that of *REG*. The similar results were found when comparing the F-measures of combined and individual scores in Table 9, except for that of *Hepatotoxicity*. Thus, overall, combined scores performed well than individual scores did not specific to particular classifier.

Table 8 Accuracies of Combined and Individual Scores by Three Classifiers

	SVM				Random Forest				Logistic Regression			
	REG	PRR	ROR	Combined	REG	PRR	ROR	Combined	REG	PRR	ROR	Combined
CV_ANG	0.593	0.712	0.712	0.729	0.507	0.557	0.557	0.650	0.502	0.652	0.652	0.626
CV_ATA	0.682	0.682	0.682	0.682	0.461	0.558	0.558	0.573	0.686	0.682	0.682	0.678
CV_CHF	0.701	0.730	0.723	0.716	0.545	0.609	0.623	0.665	0.694	0.730	0.730	0.716
CV_MI	0.600	0.655	0.655	0.664	0.473	0.473	0.473	0.545	0.518	0.655	0.655	0.673
HEP	0.620	0.601	0.600	0.617	0.481	0.463	0.469	0.533	0.612	0.616	0.616	0.613
CAN	0.710	0.710	0.710	0.710	0.575	0.572	0.568	0.624	0.710	0.710	0.710	0.710
Avg.	0.651	0.682	0.680	0.686	0.507	0.539	0.541	0.598	0.620	0.674	0.674	0.669

Table 9 F-measures of Combined and Individual Scores by Three Classifiers

	SVM				Random Forest				Logistic Regression			
	REG	PRR	ROR	Combined	REG	PRR	ROR	Combined	REG	PRR	ROR	Combined
CV_ANG	0.641	0.733	0.733	0.753	0.550	0.580	0.580	0.664	0.636	0.665	0.665	0.659
CV_ATA	0.809	0.808	0.808	0.809	0.478	0.526	0.531	0.560	0.798	0.808	0.808	0.786
CV_CHF	0.816	0.794	0.803	0.794	0.637	0.650	0.618	0.740	0.812	0.794	0.794	0.798
CV_MI	0.603	0.641	0.641	0.666	0.492	0.509	0.509	0.560	0.654	0.652	0.652	0.677
HEP	0.675	0.541	0.541	0.519	0.410	0.403	0.408	0.439	0.757	0.458	0.458	0.477
CAN	0.830	0.830	0.830	0.830	0.482	0.478	0.472	0.478	0.830	0.830	0.830	0.830
Avg.	0.729	0.725	0.726	0.729	0.508	0.524	0.520	0.574	0.748	0.701	0.701	0.705

Before we further analyze the different between combined and individual scores, it is necessary to select a classifier, which has the best performance in classification, for comparisons. Because both of Table 8 and Table 9 show that RBF-SVM has the best performance, we still chose it as the main classifier for comparison of the performance under combined and individual scores.

4.2 Marginal Improvement of Combined Model

We calculated average performance of signal detection across combined and individual scores, including average accuracy, precision, recall, and F-measure, as shown in Table 10. The relative improvement ranges from 0.3% for the F-measure of *ROR*, to an improvement of 9.5% for the recall of *REG*. Moreover, when comparing the three individual scores, *PRR* and *ROR* are at the same level of classification performance, whereas *REG* has slightly low performance.

Table 10 Marginal Improvement of Combined Model

	(1) Combined	(2) REG	(1)-(2)	(3) PRR	(1)-(3)	(4) ROR	(1)-(4)
Accuracy	0.686	0.651	5.4%	0.682	0.7%	0.680	0.9%
Precision	0.672	0.668	0.5%	0.660	1.8%	0.657	2.2%
Recall	0.480	0.439	9.5%	0.476	0.9%	0.473	1.6%
F-measure	0.729	0.729	-	0.725	0.5%	0.726	0.3%

4.3 Differences between ADRs

We also noticed that the results vary across ADRs. Table 11 displays the obvious differences in average detection performance between each ADR under combined scores. The F-measure ranges from 0.519 (*Hepatotoxicity*) to 0.830 (*Cancer*). Overall, the combined scores proposed in this research performed better in detection for two ADRs: *Angina* (All measures are > 0.7) and *Cancer* (All measures are > 0.7 except for macro-average recall).

Table 11 Average Performance of Six ADRs

	Accuracy	Precision	Recall	F-measure
CV_ANG	0.729	0.700	0.710	0.753
CV_ATA	0.682	0.682	0.333	0.809
CV_CHF	0.716	0.692	0.430	0.794
CV_MI	0.664	0.748	0.542	0.666
HEP	0.617	0.498	0.367	0.519
CAN	0.710	0.710	0.500	0.830

4.4 The Effect of Period Length

Since ADRs are usually in the form of diagnoses, and diagnoses can be classified as either acute or chronic, the length of aggregated period t may affect the performance of detection. To learn about the effect of period length, we calculated the accuracies for combined scores under different length of t (7, 30, 60, 90 and 120 days) using

RBF-SVM classifier, and the results are shown in Table 12. Interestingly, the choice of period length has no significant effect on performance, and the similar results were also found in other measures (precision, recall, and F-measure). Hence, to obtain more patient weeks, we selected shorter length (i.e., seven days) of period when aggregating visit information for each patient.

Table 12 Accuracies for Combined Scores under Different Length of Period t

Length of t	7	30	60	90	120
CV_ANG	0.729	0.745	0.729	0.743	0.731
CV_ATA	0.682	0.683	0.682	0.682	0.681
CV_CHF	0.716	0.702	0.739	0.744	0.737
CV_MI	0.664	0.700	0.700	0.709	0.718
HEP	0.617	0.607	0.615	0.610	0.603
CAN	0.710	0.710	0.710	0.710	0.710
Avg.	0.686	0.691	0.696	0.700	0.697

Chapter 5 Conclusion



5.1 Contributions

In this study, we shed light on the possibility of using health insurance claims data as data source for signal detection by developing a novel methodology, which was inspired by the precedence relationships existed among drugs and diagnoses. A series of patient weeks were built and used for calculating *REG*, *PRR*, and *ROR* score. These three scores were combined and provided as features to classification algorithms.

We also introduced a special evaluation process that carries on a classification for distinguishing serious drug-ADR pairs from other pairs. This process may provide a new direction for initially screening out serious drug-ADR pairs.

Through different analyses we demonstrated that the proposed combined score led to an improvement of signal detection accuracy over applying each of score independently. We also showed that the results varied across ADRs but not the length of aggregate period.

5.2 Managerial Implication

Being routinely collected and covering over 99% of the population in Taiwan (Bureau of National Health Insurance, Department of Health, Executive Yuan, Taiwan, 2012), health insurance claims data involve more drugs and diagnoses than existed reporting systems. Therefore, if these claims data can be used for ADR signaling and detect possible ADRs for drugs in early post-marketing phase, it is expected to significantly reduce the number of ADR reports and then save time and cost to medical staff for reporting ADRs.

5.3 Limitations and Future Work

Our study assumed that drug usage may have the chance to cause ADRs one week later. In practice, however, the time from using a drug to an attack of an adverse event may be much longer. Although our results showed that there was no significant effect between different lengths of aggregated period in the detection performance, time factor is still an important issue when detecting ADRs in health databases. For instance, we can consider two nonadjacent patient weeks when building the precedence relationship between drugs and diagnoses to represent the deferred effect of drugs.

Moreover, our dataset covers near 1.5 million drug-diagnosis pairs (the combination of 1,125 diagnoses and 1,326 drugs); however, ADRs form a relatively small part of all the diagnoses in NHIRD. This data imbalance problem may bring difficulties when screening drug-ADR pairs in all drug-diagnosis pairs.

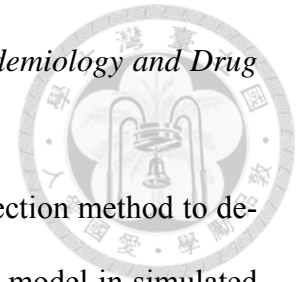
Finally, detection performance may be further improved by bringing more properties, e.g., dosage of the drug, to the aggregated patient weeks. Other scores introduced in Chapter 2, e.g., information component, may be added to the combined detecting scores. However, we should consider the impact of feature expansion on time efficiency (in our study, it took less than four hours to calculate scores for 7.5 million patient weeks by a computer equipped with an Intel Core-i7 3.2Ghz CPU).

References



- Bate, A., & Evans, S. J. W. (2009). Quantitative signal detection using spontaneous ADR reporting. *Pharmacoepidemiology and Drug Safety*, 18(6), 427–436. doi:10.1002/pds.1742
- Bate, A., Lindquist, M., Edwards, I. R., & Orre, R. (2002). A data mining approach for signal detection and analysis. *Drug Safety: An International Journal of Medical Toxicology and Drug Experience*, 25(6), 393–397.
- Bates, D. W. (1995). Incidence of Adverse Drug Events and Potential Adverse Drug Events: Implications for Prevention. *JAMA*, 274(1), 29. doi:10.1001/jama.1995.03530010043033
- Bright, T. J., Wong, A., Dhurjati, R., Bristow, E., Bastian, L., Coeytaux, R. R., ... Lobbach, D. (2012). Effect of clinical decision-support systems: a systematic review. *Annals of Internal Medicine*, 157(1), 29–43. doi:10.7326/0003-4819-157-1-201207030-00450
- Bureau of National Health Insurance, Department of Health, Executive Yuan, Taiwan. (2012, May). Universal Health Coverage in Taiwan. Retrieved from http://www.nhi.gov.tw/Resource/webdata/21717_1_20120808UniversalHealthCoverage.pdf
- Choi, C. A., Chang, M. J., Choi, H. D., Chung, W.-Y., & Shin, W. G. (2013). Application of a drug-interaction detection method to the Korean National Health Insurance claims database. *Regulatory Toxicology and Pharmacology: RTP*, 67(2), 294–298. doi:10.1016/j.yrtph.2013.08.009
- Coloma, P. M., Trifirò, G., Schuemie, M. J., Gini, R., Herings, R., Hippisley-Cox, J., ... EU-ADR Consortium. (2012). Electronic healthcare databases for active drug

safety surveillance: is there enough leverage? *Pharmacoepidemiology and Drug Safety*, 21(6), 611–621. doi:10.1002/pds.3197



Cornelius, V. R., Sauzet, O., & Evans, S. J. W. (2012). A signal detection method to detect adverse drug reactions using a parametric time-to-event model in simulated cohort data. *Drug Safety: An International Journal of Medical Toxicology and Drug Experience*, 35(7), 599–610.

doi:10.2165/11599740-000000000-00000

Deshpande, G., Gogolak, V., & Smith, S. W. (2010). Data Mining in Drug Safety: Review of Published Threshold Criteria for Defining Signals of Disproportionate Reporting. *Pharmaceutical Medicine*, 24(1), 37–43. doi:10.1007/BF03256796

Dumouchel, W. (1999). Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System. *The American Statistician*, 53(3), 177–190. doi:10.1080/00031305.1999.10474456

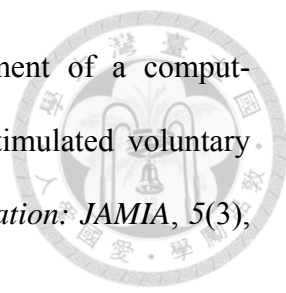
Hacker, M. P., Messer, W. S., & Bachmann, K. A. (2009). *Pharmacology principles and practice*. Amsterdam; Boston: Academic Press/Elsevier. Retrieved from <http://public.eblib.com/EBLPublic/PublicView.do?ptiID=452816>

Harpaz, R., Chase, H. S., & Friedman, C. (2010). Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics*, 11(Suppl 9), S7. doi:10.1186/1471-2105-11-S9-S7

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

doi:10.1136/amiajnl-2012-000930

Jha, A. K., Kuperman, G. J., Teich, J. M., Leape, L., Shea, B., Rittenberg, E., ... Bates,

- 
- D. W. (1998). Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. *Journal of the American Medical Informatics Association: JAMIA*, 5(3), 305–314.
- Jin, H., Chen, J., He, H., Kelman, C., McAullay, D., & O’Keefe, C. M. (2010). Signaling Potential Adverse Drug Reactions from Administrative Health Databases. *IEEE Transactions on Knowledge and Data Engineering*, 22(6), 839–853. doi:10.1109/TKDE.2009.212
- Johansson, S., Wallander, M.-A., de Abajo, F. J., & García Rodríguez, L. A. (2010). Prospective drug safety monitoring using the UK primary-care General Practice Research Database: theoretical framework, feasibility analysis and extrapolation to future scenarios. *Drug Safety: An International Journal of Medical Toxicology and Drug Experience*, 33(3), 223–232. doi:10.2165/11319010-000000000-00000
- Johnson, J. A. (1995). Drug-Related Morbidity and Mortality: A Cost-of-Illness Model. *Archives of Internal Medicine*, 155(18), 1949. doi:10.1001/archinte.1995.00430180043006
- Kubota, K., Koide, D., & Hirai, T. (2004). Comparison of data mining methodologies using Japanese spontaneous reports. *Pharmacoepidemiology and Drug Safety*, 13(6), 387–394. doi:10.1002/pds.964
- Kuperman, G. J., Bobb, A., Payne, T. H., Avery, A. J., Gandhi, T. K., Burns, G., ... Bates, D. W. (2007). Medication-related clinical decision support in computerized provider order entry systems: a review. *Journal of the American Medical Informatics Association: JAMIA*, 14(1), 29–40. doi:10.1197/jamia.M2170
- Lazarou, J., Pomeranz, B. H., & Corey, P. N. (1998). Incidence of adverse drug reac-

- tions in hospitalized patients: a meta-analysis of prospective studies. *JAMA: The Journal of the American Medical Association*, 279(15), 1200–1205.
- Miniño, A. M., Murphy, S. L., Xu, J., & Kochanek, K. D. (2011). Deaths: final data for 2008. *National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, 59(10), 1–126.
- National Health Insurance Administration, Ministry of Health and Welfare. (2014, January 3). 藥品代碼與 ATC 碼對照. Retrieved from [http://www.nhi.gov.tw/Resource/webdata/13733_1_藥品代碼與 ATC 碼對照-10212\(上網\).xls](http://www.nhi.gov.tw/Resource/webdata/13733_1_藥品代碼與 ATC 碼對照-10212(上網).xls)
- Norén, G. N., Sundberg, R., Bate, A., & Edwards, I. R. (2008). A statistical methodology for drug-drug interaction surveillance. *Statistics in Medicine*, 27(16), 3057–3070. doi:10.1002/sim.3247
- Orre, R., Lansner, A., Bate, A., & Lindquist, M. (2000). Bayesian neural networks with confidence estimations applied to data mining. *Computational Statistics & Data Analysis*, 34(4), 473–493. doi:10.1016/S0167-9473(99)00114-0
- 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(6), 598–607. doi:10.1002/pds.2139
- Pharmaceutical data mining: approaches and applications for drug discovery*. (2010). Hoboken, N.J: Wiley.
- Sauzet, O., Carvajal, A., Escudero, A., Molokhia, M., & Cornelius, V. R. (2013). Illustration of the weibull shape parameter signal detection tool using electronic

healthcare record data. *Drug Safety: An International Journal of Medical Toxicology and Drug Experience*, 36(10), 995–1006.

doi:10.1007/s40264-013-0061-7



Uppsala Monitoring Centre. (2013, April 22). VigiBase. Retrieved from <http://www.who-umc.org/DynPage.aspx?id=98082&mn1=7347&mn2=7252&mn3=7322&mn4=7326>

World Health Organization, & WHO Collaborating Centre for International Drug Monitoring. (2002). *The importance of pharmacovigilance*. [Geneva]: World Health Organization : Uppsala Monitoring Centre, WHO Collaborating Centre for International Drug Monitoring.

蔡雅婷, 陳文雯, & 蔡翠敏. (2014). 102 年度藥品不良品通報系統之案件分析. *Drug Safety Newsletter*, (45), 18–27.