

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

資料探勘技術於嚴重藥物不良反應之探測研究

On Detecting Serious Adverse Drug Reaction

with Data Mining Techniques

黃嫩雅

Samantha Hwang

指導教授: 曹承礎 博士

Advisor: Seng-Cho Chou, Ph. D.

中華民國106年7月

July, 2017

國立臺灣大學(碩、博)士學位論文

口試委員會審定書



(題目: On Detecting Serious Adverse Drug Reaction with Data Mining Techniques)

本論文係 黃嫩雅 君 (學號 R04725038) 在國立臺灣大學資訊管理學系、所完成之 (博、碩) 士學位論文, 於民國 106 年 7 月 3 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:

曹月如

盧信鈞

謝志誠

所 長:

蔡益坤

誌謝



碩士生活兩年，非常謝謝曹承礎老師的指導。從一開始的懵懂無知到現在完成碩士學位的時刻，老師都會全力地給予支持與鼓勵。在學業上，從一開始思考題目方向、經過中間研究的重重關卡、到最後的完成品，老師都會鍥而不捨的提點我們新的方向。在生活上，如果遇上任何問題，老師也都不吝嗇的給予幫助。老師也很注重學生的實務經驗。在碩士一年級時，老師會利用身邊的資源，來讓我們跟公司接觸，了解公司在實作資料探勘技術的流程，以讓我們與產業接軌。碩士兩年來，收穫非常的豐富。

除了感謝曹老師外，也特別感謝謝冠雄老師以及盧信銘老師。在口試時，給我許多寶貴的意見以及想法，讓這篇論文更具貢獻價值。而在研究的過程中，非常感謝林煥博醫師的輔導。林醫師給予我們許多醫學上的知識，也幫助我們取得難能可貴的資源來幫助我們做研究。林醫師向來助人為樂，能認識林醫師真的非常的幸運。

另外，在兩年的碩士生活中，謝謝Tina、中彥、小康以及顯鈞一直以來的照顧以及支持，你們都是實驗室不可或缺的角色。謝謝阿波不管在任何時候都會在旁給我鼓勵，非常開心能在台大認識你，讓我的碩士生活增添許多美好回憶。謝謝雅歆以及芝伊在學業上以及生活上，有任何需要幫忙或是出遊玩樂都會揪來揪去，覺得非常溫暖。

最後，非常感謝我的爸媽以及弟弟。如果沒有他們，就不會有樂觀的嫩雅。一直以來的全力支持與鼓勵，讓我能順利的完成學業。

摘要



現今，有越來越多吃了食品與藥物管理署認證過的藥物後而得到未預期的嚴重藥物不良反應的案例。嚴重的藥物不良反應狀況有包含死亡、生命危急、需要住院或是延長住院時間、長期或顯著的身體殘障、致畸胎以及其他任何會導致前面五種狀況的疾病。醫院或是病人如果有發現這些嚴重不良反應可以向通報系統通報，而這些藥物就會被重新檢驗。這篇研究主要是希望能夠偵測嚴重藥物不良反應來改善目前的效率以及低估問題。為了資料的完整性，這篇研究選擇使用台灣衛生署健保局的健保資料庫來當我們的資料庫。我們希望能利用關連規則從健保資料庫中找出藥物以及未預期的嚴重藥物不良反應之間的關係。建立的規則會以 leverage 以及 unexlev 值來做篩選。在結果上，我們發現有兩種藥物，cisapride 以及 terfenadine，所導致的嚴重不良反應可以比通報系統更早被我們偵測到。另外，從丹麥藥品局的報告中，高頻率會導致嚴重藥物不良反應的藥物也在我們篩出來的規則中比較前面的名次。這篇研究所篩選出來的規則可以經由專家驗證後，提早警示食品與藥物管理署來對這些藥物重新做檢驗。

關鍵字：全民健保資料庫、關聯規則、嚴重藥物不良反應、丹麥藥品局

Abstract



Nowadays, more and more people occurs unexpected serious reaction after taking an FDA-approved drug. Reactions such as death, life-threatening, requires inpatient hospitalization or prolongation of existing hospitalization, persistent or significant disability/incapacity, a congenital anomaly/birth defect, or other situations will be reported to the hospitals and the drugs may be examined again. This study intends to discover the potential unexpected serious ADRs automatically by using the data mining techniques in order to improve the efficiency of detecting and to avoid the under-reporting biases. For the completeness of every patients' records, we chose the NHIRD as our database. We want to find the strong links between the drugs, which the patients took, and the unexpected serious ADRs, which the patients suffer after taking the drug, by the association rules. Rules would be obtained and chose according to the leverage and unexlev threshold. We found that the serious ADRs of cisapride and terfenadine can be detected earlier than reporting system. The high frequency of drugs that would cause serious ADRs listed by the Danish Medicines Agency's network were found in a high rank. Experts may examine the rules we selected and alarm the FDA for these highlighted relationships.

Key Words: National Health Insurance Research Database, association rules, serious

adverse drug reaction, Danish Medicines Agency's network

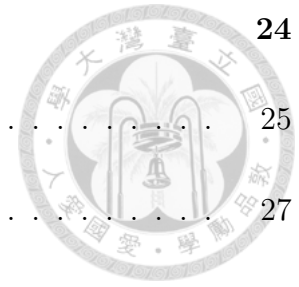




Contents

1 Introduction	1
1.1 Motivation	1
1.2 Research objectives	3
1.3 Research structure	4
2 Literature Review	6
2.1 National Health Insurance Research Database	6
2.2 Medical Literature	8
2.3 Technique Literature	9
3 Method	11
3.1 Concept of Method	11
3.2 Data Collection	18
3.3 Data Preparation	20
3.4 Data Analysis	23

4	Experimental Results	24
4.1	Validation	25
4.2	Overall Results	27
4.3	Description of Potential Rules with Serious Diseases	30
4.4	Results of Potential Rules with Serious Diseases	38
5	Conclusion	44
5.1	Contribution	44
5.2	Limitation and Future Work	45
	Bibliography	47

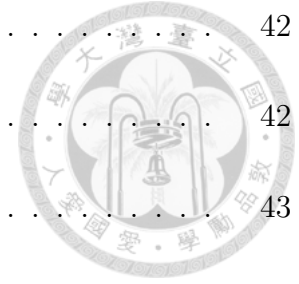




List of Figures

1.1	Structure of this study	5
3.1	Figure of filtering out severe diseases patients	12
3.2	Getting potential drug list	13
3.3	Dataset extracted from NHIRD	14
3.4	Concept of TAR	15
3.5	Concept of UTAR	17
3.6	People who got disease between 1999 to 2003	21
3.7	People who got serious disease between 1999 to 2003	22
4.1	Amount of Records in each year	32
4.2	Amount of Patients in each year	33
4.3	Distribution of Sex Type	34
4.4	Distribution of Age	38
4.5	Distribution of taken drugs	39
4.6	Amount of Records of Group "C" in each year	40

4.7	Amount of Patients of Group “C” in each year	42
4.8	Distribution of Sex Type in Group “C”	42
4.9	Distribution of Age Group in Group “C”	43





List of Tables

4.1	Corresponding names of cisapride, terfenadine, and QT interval prolongation	25
4.2	Validated rules of cisapride and terfenadine in the selected rules (N = 350,035,533)	26
4.3	List of drugs that involve serious ADRs with corresponding ICD-9-CM code	28
4.4	Validated rules appear in the selected rules (N = 350,035,533)	29
4.5	Rules of top 15 base on RR (N = 350,035,533)	31
4.6	The proportion of sex type	33
4.7	Age Group	35
4.8	The amount of people in records and population	35
4.9	ATC-code Classification Principle	36
4.10	Amount of records taking drugs classified by ATC Group	37
4.11	Rules of top 15 base on RR in serious diseases (N = 350,035,533)	41

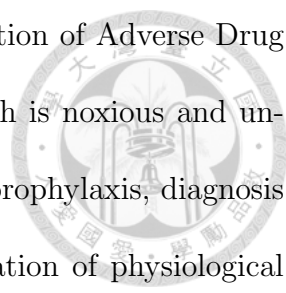


Chapter 1

Introduction

1.1 Motivation

In USA, when there are new drugs that intend to be launched, they need to be strictly examined and be approved by the new drug's clinical trials of the U.S. Food and Drug Administration(USFDA). The USFDA separates the new drug investigation into four phases. They look for healthy volunteers to define the most frequent side effects of the new drug in phase one. They search for patients who have certain disease to see whether the new drug will work in these patients or not in phase two. After they get the initial data in phase one and phase two, they will try to let patients who are in different populations to take different dosages of the new drug. This decision will base on their condition and will try different experiments by combining them with other drugs for gathering more data in phase three. Finally, if the new drug does not cause severe side effects, the USFDA will approve the new drug and will keep monitoring after the new drug has been launched in phase four.

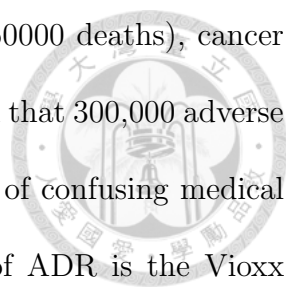


According to the World Health Organization(WHO), the definition of Adverse Drug Reaction(ADR) in 1972 is a response to a medicinal product which is noxious and unintended and which occurs at doses normally used in man for the prophylaxis, diagnosis or therapy of disease or for the restoration, correction or modification of physiological function. An adverse drug reaction, contrary to an adverse event, is characterized by the suspicion of a causal relationship between the medicine and the occurrence, i.e. judged as being at least possibly related to treatment by the reporting or a reviewing health professional.¹ ADR could probably occur in phase four of the new drug investigation. While on the basis of the definition above, the USFDA divided the ADR into several groups. What this thesis emphasize is the unexpected serious adverse drug reactions. An unexpected serious adverse drug reaction is any untoward medical occurrence that at any dose, which the nature or severity of the reaction is not consistent with information in the relevant source documents. This medical occurrence results in death, life-threatening (which refers to an event in which the patient was at risk of death at the time of the event), requires inpatient hospitalization or prolongation of existing hospitalization, persistent or significant disability/incapacity, a congenital anomaly/birth defect, or other situations (such as important medical events that may not be immediately life-threatening or result in death or hospitalization but may jeopardize the patient or may require intervention to prevent one of the other outcomes listed in the former).²

As mentioned by Katzung (2015), ADR is said to be the fourth leading cause of death. The University of Toronto estimated that ADRs could have more than 100,000

¹http://www.who.int/medicines/areas/coordination/English_Glossary.pdf

²<http://www.fda.gov/downloads/Drugs/.../Guidances/ucm073087.pdf>



deaths in the USA each year, coming after heart disease (about 750000 deaths), cancer (530000), and stroke (150000). (Bonn, 2005) However, FDA asserted that 300,000 adverse events occurred in hospitals may be preventable, many as a result of confusing medical information or lack of information in 2015. A well-known case of ADR is the Vioxx occurred in 2004. Vioxx was approved for use against arthritis by the USFDA and had been on the market since May,1999. There were millions of arthritis taking this drug. The sales volume was up to \$2.5 billion, which is equivalent to NT\$80.6 billion, in 2003. However, a study indicated that taking Vioxx for more than 18 months could increase the patients' risks for thrombotic cardiovascular events. (Bresalier et al., 2005) The Merck Sharp & Dohme (MSD), who manufactured the Vioxx, pulled the drug from market after they found that this drug might cause for more than 18 risks for heart attacks and strokes in some studies. The USFDA testified and proofed about the studies, which led to withdrawing the drug in November, 2004. This situation has caused dominant focus in the world.

1.2 Research objectives

For improving the efficiency of detecting adverse drug reaction, there are more and more researchers combining ADR with different technology methods in other countries. In Taiwan, the most prominent method is using the reporting systems. Patients may spontaneously report the potential ADR via the Internet or post to the Taiwan Drug Relief Foundation(TDRF). TDRF was created and dedicated to carry out Drug Relief Law. The Foundation served to receive patients' applications for drug relief, educates people

on drug safety, collects relief fund, releases subsidy, and establishes databases for Pharmacovigilance.³ However, this reporting system could suffer from under-reporting biases for not all of the patients will report the potential ADR they found. Even if the TDRF finally found that there were links between the drug and the serious unexpected ADR, the losses caused by the reaction may be severe already.

This study intends to discover the potential unexpected serious ADRs in Taiwan by using the data mining techniques in order to improve the efficiency of detecting and to avoid the under-reporting biases. With the integrity of the Taiwan's National Health Insurance Research Database(NHIRD), we can get almost all of the Taiwanese patients' medical records from 1997 to 2012 by this database. We want to find the strong links between the drugs, which the patients took, and the unexpected serious ADRs, which the patients suffer after taking the drug, by association rules. Rules would be obtained and be chosen according to the rank we had made.

1.3 Research structure

This thesis will be organized as shown in Figure 1.1 . We introduced our motivation and objective in the front, and will review the medical and technique literatures in the next chapter. The method is proposed to process the data from NHIRD in Chapter 3. The experimental results and the evaluation of this model are presented and discussed in Chapter 4. At last, this study is concluded in Chapter 5, with the limitation of the research methods and the future works of this study.

³http://www.tdrf.org.tw/en/01_about/abo_01_list.asp

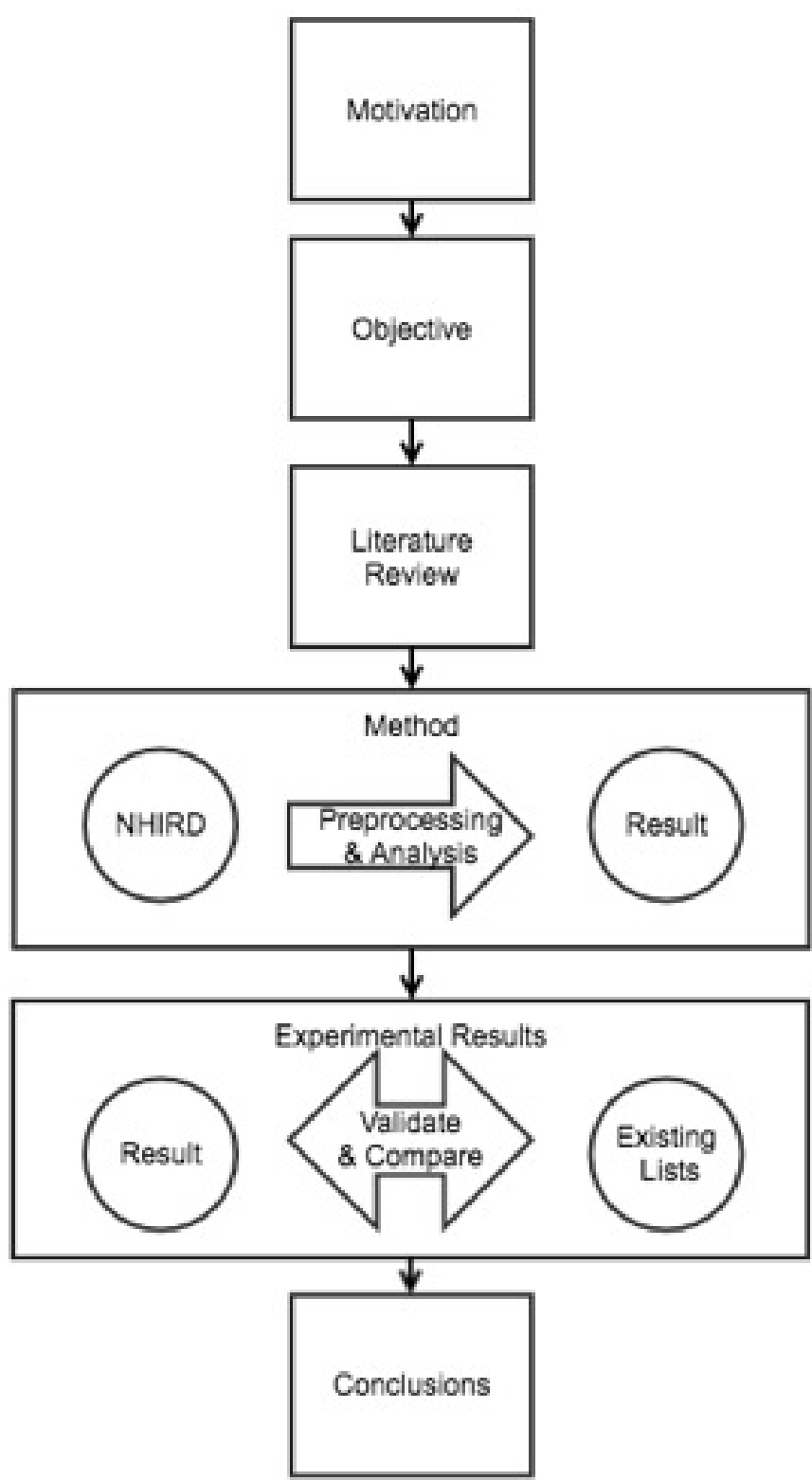


Figure 1.1: Structure of this study



Chapter 2

Literature Review

2.1 National Health Insurance Research Database

Taiwan launched a National Health Insurance(NHI) program on March 1, 1995. There was around 99.9% of Taiwan's population were enrolled in this program, and 93% of the hospitals and clinics in Taiwan are NHI-contracted until now (National Health Insurance Administration, 2015). The database of this program involves the data from the beginning of the patients' appointment information to the end of the patients' medical records. Maintained by the National Health Research Institutes(NHRI) in Taiwan, this database provides the information for researchers to do medicine-related study and for merchants to do more value-added services, respectively.¹

The NHIRD has massive amounts of data. It will take a lot of time to handle all of the data by the computers, and will be unfavorable to privacy protection. The Longitudinal

¹<http://nhird.nhri.org.tw/en/index.html>

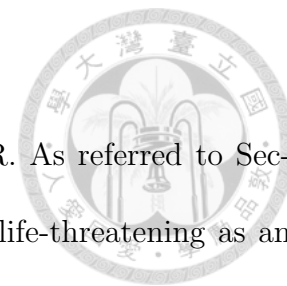
Health Insurance Database(LHID) is a database that provides representative sampled data for researchers to study. There are three issues of LHID currently: LHID2010, LHID2005, and LHID2000. In this study, the LHID2010 database is chosen to be our dataset. According to the NHRI, LHID2010 is composed by the registration and claimed data of 1,000,000 individuals, who are random samples of the NHI program's beneficiary during the period of January 1st, 2010 to December 31st, 2010.

The NHRI grouped these 1,000,000 individuals by encrypted social security numbers into 25 groups. Each group has 40,000 individuals and is connected with the NHIRD to access all of the medical records of these individuals from 1996 to 2010. The LHID2010 was updated every year to add the new medical records, appended in the new year, of these 1,000,000 individuals. The connected data in the NHIRD included Ambulatory care expenditures by visits(CD), Details of ambulatory care orders(OO), Details of inpatient orders(DO), Inpatient expenditures by admissions(DD), Expenditures for prescriptions dispensed at contracted pharmacies(GD), Details of prescriptions dispensed at contracted pharmacies(GO), and Registry for beneficiaries(ID).²

In order to understand the data efficiently, the NHRI has provided the codebook of the NHIRD. There are two parts of this codebook: the data description and the code description. The data description describes how the data looks like in the database, including the column names, the corresponding data type, the length of the data, the start and end position of the data, and the definition. On the other hand, the code description shows what the data, which performed as a code number, represents in words(National Institutes of Health, 2014).

²http://nhird.nhri.org.tw/en/Data_Subsets.html#S1

2.2 Medical Literature



In this study, we only focus on the drugs that cause serious ADR. As referred to Section 1.1, there are six cases that belongs to serious ADRs. Take life-threatening as an example, we've known that drugs that contain cisapride or terfenadine may cause QT interval prolongation. Cisapride is a gastrointestinal prokinetic agent for the treatment of gastroesophageal reflux, functional dyspepsia, gastroparesis, chronic constipation and irritable bowel syndrome(Wang et al., 2001). It has been found that this drug may lead to inappropriate lengthening of the QT interval and induction of major cardiac rhythm disturbances. This situation not only happens in patients taking high doses of cisapride, but also those who receive clinically recommended doses. On the other hand, terfenadine is another drug that was used to treat hay fever(Waller, 2009). This treatment has the occurrence of prolonged QT interval leading to ventricular tachycardia of torsades de points (Shaikh, 2000).

In Taiwan, cisapride was used until 2004, while terfenadine reached till 2005. According to the local ADR reporting system, there were 6 patients, who took cisapride, result in QT interval prolongation. Their age range from forty to seventy years old and one of them died because of this adverse drug event. There were also some cases show that terfenadine leads to cardiac arrhythmia and died at last. The database we have chosen includes the range they were been used. We expect that the rules of the two drugs can be detected for verifying the rank we made is accurate.

Moore et al. (1998) claimed that ADRs usually take place on older women. They took more drugs and stayed longer in hospitals. This also led to having more ADRs. This

study estimated that 70% of adverse drug reactions may be preventable, while serious adverse drug reactions occupied a half among them. There will be a yearly saving of over £50,000 if the avoidable serious ADRs were prevented. These savings may turn out to treat other patients and moreover improve the hospital or healthcare system efficacy and productivity. In this case, detecting serious ADRs seems to be much more significant.

In Taiwan, Chan et al. (2008) mentioned that the most severe dermatologic ADR was Stevens-Johnson syndrome, while Hematologic ADRs followed behind. The experimental results Chan et al. (2008) proposed marked that 84% of ADRs were considered predictable and there was a cost of US\$3489.00 to treat those with ADRs in each hospital. While the ADR reporting rate were claimed low in the study, we want to improve this process and allocate the resources to those who really need.

2.3 Technique Literature

We may see that Hsieh (2014) proposed a novel EHR-based drug safety signal detection method on the basis of the learning with the NHIRD to rank approach in Taiwan. This method significantly outperforms the benchmarks. However, we want to find out the unexpected serious ADRs. We may focus more on how to detect the unexpected.

Chazard et al. (2011) have discussed about how to generate adverse drug events by the application of data mining. The data they used were extracted from electronic health records from six different hospitals. They mainly used the aggregation engines to find out the rules that what kinds of ADE might be conformed from the drugs and the conditions. However, we only pay attention on the rules between drug and the ADR, instead of

including the age ,the weight, and other conditions.

In Australia, McAullay et al. (2009) proposed a method, UTARs, to find the strong links via association rules. They consider the conditions only by the drugs and the disease. Furthermore, the UTARs is expected to find out the unexpected patterns in the database. They will count the events that the disease first happens after the taken drug. This method can also be helpful to avoid the problem of setting thresholds. It is known that getting the right rules by setting the right threshold is difficult.

Reps et al. (2013) also made use of method above on The Health Improvement Network (THIN) database. The study mentioned that the algorithm is good to find ADRs when the number of patients prescribed the drug is low. So, Reps et al. (2013) suggested this method may be the optimal algorithm to apply when a drug is newly marketed.

In this study, we will make use of the method proposed by McAullay et al. (2009). As mentioned in Section 1.2, we want to find out what drugs might cause serious ADRs. The concept of how to detect serious ADR by using the proposed method by McAullay et al. (2009) will be discussed in Chapter 3.



Chapter 3

Method

3.1 Concept of Method

As McAullay et al. (2009) mentioned, finding the patterns of unexpected and infrequent ADRs is difficult for three reasons:

1. Drugs are strictly screened before launching.
2. Launched drugs are still strictly screened for concerning any adverse event happens.
3. A drug is strong associated with certain diseases that is purposely prescribed for treatment.

For solving the problems, we referred to the model proposed by McAullay et al. (2009) and split this experiment into three steps, including the initializing step.

Concerned merely about the serious ADRs, we filtered out the patients who have severe diseases to simply find out drugs that have potential serious ADRs in the initializing

step. As shown in Figure 3.1, $C1$, $C2$, $C3$ are the patients who got severe diseases in the NHIRD. After filtering out the subset of NHIRD, we defined a variable T to indicate the

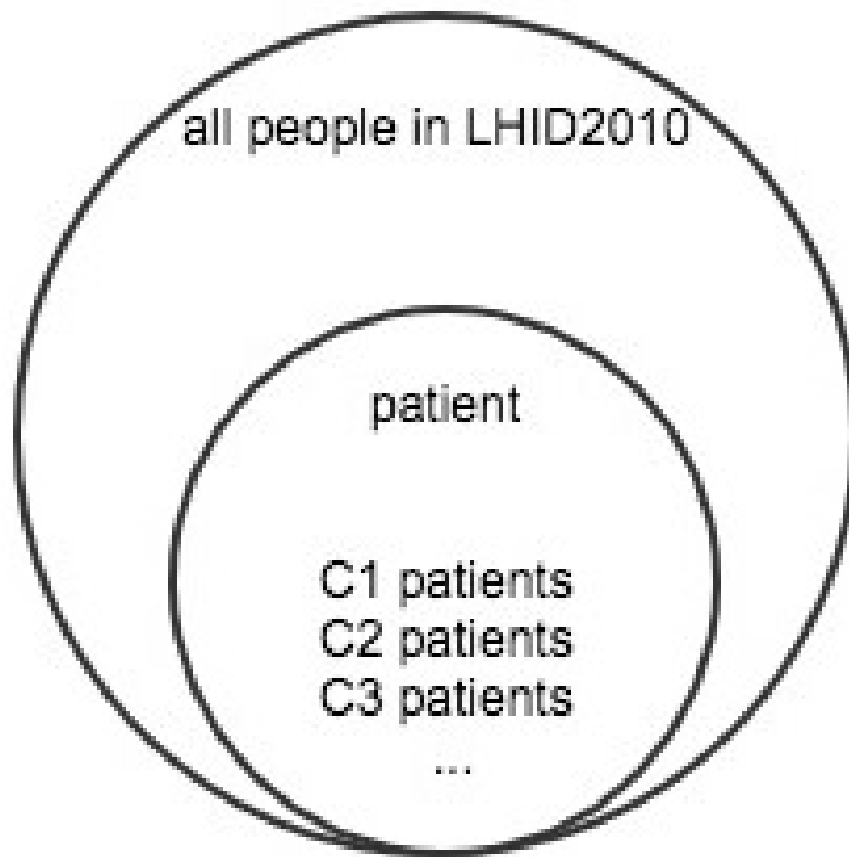


Figure 3.1: Figure of filtering out severe diseases patients

time period of drug effect. In this study, we considered $T = 6$ for drugs can effect on our body at most six months. Potential drugs, $A1$, $A2$, $A3$, $A4$ were received by all the drugs patients took in the period of T months before first diagnosed the severe disease (shown in Figure 3.2). A patient might have $C1$, $C2$, $C3$ together. The main purpose of this step is to reduce all of the drugs in the database to a smaller range that certain drugs that might cause serious ADRs. For surely completeness of obtaining potential drugs, we observe the patient three times respectively depending on the kinds of severe disease

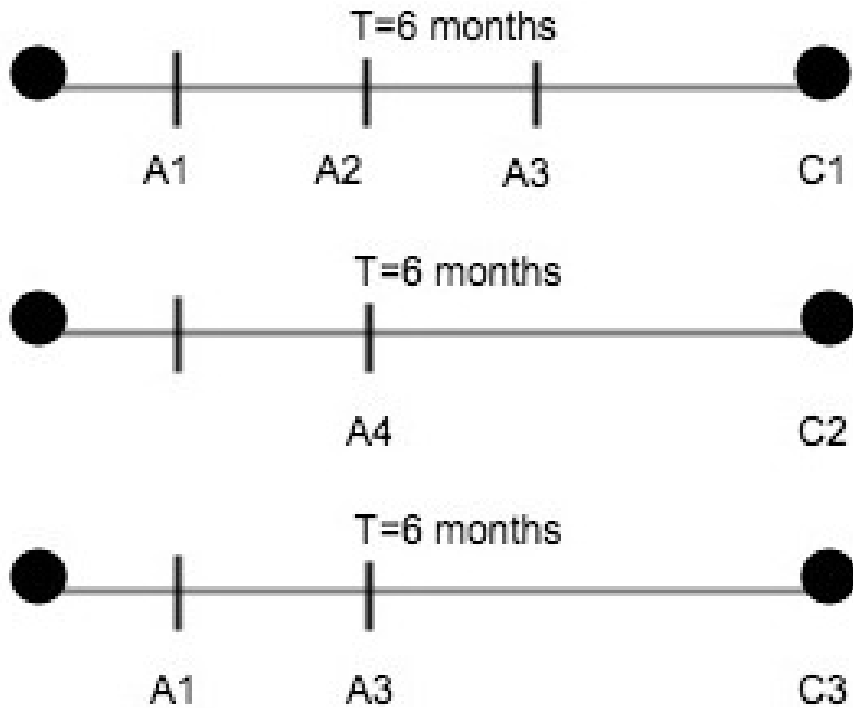


Figure 3.2: Getting potential drug list

he got in the above example. We use this potential drug list to find out those who took these drugs and establish our data set (shown in Figure 3.3). The records that appears to be the potential drugs has been first taken by the patients will be extracted for the next step.

In the next step, we adopted the Temporal Association Rule(TAR). The patterns like *the antecedent A* followed by *the consequent C* in a time window length of T , denoted by $A \xrightarrow{T} C$, are extended from association rules. Potential drugs that lead to serious ADRs are what we concerned. Therefore, we take A as a first taken potential drug and C as a disease happened after taking the drug in a time period T . The list of potential drug A has been acquired in the initialization step. In this model, the *event – oriented data preparation* is proposed. As concerned for the one-by-one relationship between a drug



Figure 3.3: Dataset extracted from NHIRD

and a disease, we will split the events to many sub-events. Taken Figure 3.4 as example, there are three events and seven sub-events, e.g. $[A1, C4]$ and $[A1, C1]$ are two sub-events in the first event, in this case. These events happen to be the first time users taken the drug. Hence, the former two events must come from two different users. Potential drug $A1$ users have subsequences of $\{C1, C2, C4\}$ and $\{C1, C5\}$, while potential drug $A2$ user has subsequence of $\{C1, C2\}$. The sequence in the subsequence will not be discussed in this study for drugs. We mainly discuss about the existence pattern for simplicity. In this case, we will produce six rules as following:

$$A1 \rightarrow C1$$

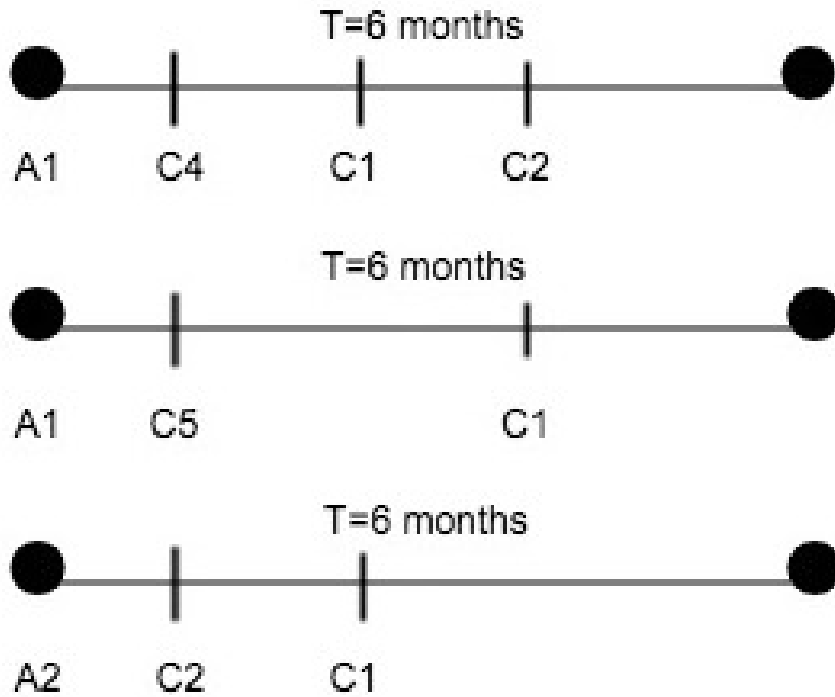


Figure 3.4: Concept of TAR

$$A1 \rightarrow C2$$

$$A1 \rightarrow C4$$

$$A1 \rightarrow C5$$

$$A2 \rightarrow C1$$

$$A2 \rightarrow C2.$$

After generating the rules, we will do some statistics to filter out the unimportant rules. The measure of *support*, *confidence*, and *leverage* are used in this study. The *support*, denoted by

$$supp(A \xrightarrow{T} C), \quad (3.1)$$

indicates how frequently A occurs before C in given time period T . In Figure 3.4, let's

take $A1 \rightarrow C1$ as an example. The support of this rule is

$$\text{supp}(A1 \rightarrow C1) = \frac{2}{7}.$$

The *confidence*,

$$\text{conf}(A \xrightarrow{T} C) = \frac{\text{supp}(A \xrightarrow{T} C)}{\text{supp}(A \xrightarrow{T})}, \quad (3.2)$$

represents the proportion of rules containing A . We get the value,

$$\text{conf}(A1 \rightarrow C1) = \frac{2}{5},$$

in Figure 3.4. The measurement *leverage* shows us whether the *antecedent* A influences the *consequent* C , denoted by

$$\text{leverage}(A \xrightarrow{T} C) = \text{supp}(A \xrightarrow{T} C) - \text{supp}(A \xrightarrow{T}) \times \text{supp}(C). \quad (3.3)$$

As shown in Figure 3.4, we can see that the *leverage* of the rule $A1 \rightarrow C1$ is

$$\text{leverage}(A1 \rightarrow C1) = \frac{2}{7} - \frac{5}{7} \times \frac{3}{7} = -\frac{1}{49}.$$

The range of *leverage* is between $[-1, 1]$, while zero indicates that these two variables are independent. The generated rules will be filtered by the given thresholds of these three measures. However, McAullay et al. (2009) mentioned that setting the right thresholds to get the right rules is difficult. Hence, we move on to the last step.

The Unexpected Temporal Association Rule(UTAR) is adopted in the last step. The purpose of this model is to strengthen the unexpected rules to a higher rank. Therefore, we may denote this method by $A \xrightarrow{T} C$ as a meaning of the *consequent* C happens unexpectedly after the *antecedent* A with a time period T . By conducting this experiment, we might want to know which kinds of diseases did the patients got before taking



the potential drugs. We need a reference time period T' to obtain the diseases we want to prune. In each of the events we gathered in TAR, we search for the sequences that happens before A in a time period T' . We set $T' = T = 6$ for simplicity. The sequence found in the reference time would be excluded in the subsequence of A if there were any match. This may prune the expected ADRs and weaken the common diseases. A concept of UTAR is shown in Figure 3.5. The subsequence of $A1$ and is $\{C1, C2, C4\}$ and $\{C2\}$,

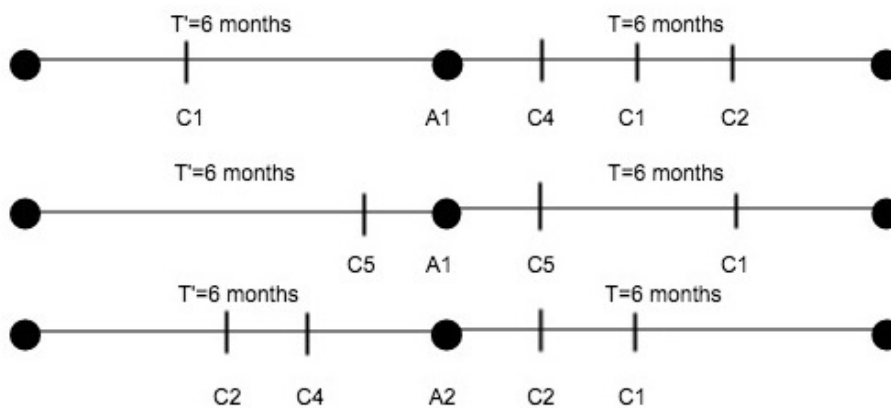


Figure 3.5: Concept of UTAR

while $A2$ is $\{C1\}$ in this case. We may acquire the rules as following:

$$A1 \leftrightarrow C4$$

$$A1 \leftrightarrow C2.$$

$$A1 \leftrightarrow C1$$

$$A2 \leftrightarrow C1.$$

As similar to TAR, we count the *support* and the *confidence* which is represented by

$$supp(A \xrightarrow{T} C), \tag{3.4}$$

and

$$conf(A \overset{T}{\leftrightarrow} C) = \frac{supp(A \overset{T}{\leftrightarrow} C)}{supp(A \overset{T}{\rightarrow})}, \quad (3.5)$$

respectively. Unlike the *leverage* in TAR, it is named *unexlev* and denoted by

$$unexlev(A \overset{T}{\leftrightarrow} C) = supp(A \overset{T}{\leftrightarrow} C) - supp(A \overset{T}{\rightarrow}) \times supp(\overset{T}{\leftrightarrow} C). \quad (3.6)$$

The *unexlev* indicates how much degree of A influences unexpected C , while $supp(\overset{T}{\leftrightarrow} C)$ is the proportion of rules that contain unexpected C . Similar to *leverage*, the range is between $[-1,1]$, while zero indicates the two variables are independent. In the sub-events that have different *antecedent*, the different ones will be considered as containing unexpected C if they have *consequent* C . Take $A1 \leftrightarrow C1$ as an example from Figure 3.5, the *support* is

$$supp(A1 \leftrightarrow C1) = \frac{1}{7},$$

the *confidence* is

$$conf(A1 \leftrightarrow C1) = \frac{1}{5},$$

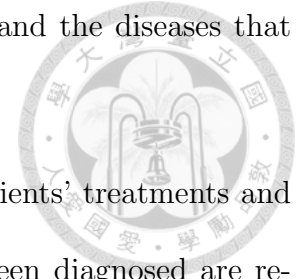
and the *unexlev* is

$$unexlev(A1 \leftrightarrow C1) = \frac{1}{7} - \frac{5}{7} \times \frac{2}{7} = -\frac{3}{49}.$$

3.2 Data Collection

The data in the LHID2010 is used in this study. As mentioned previously in Section 2.1, these accessed data were the 1,000,000 individuals randomly extracted from the NHIRD. Collected from 1997 to 2012, the Ambulatory care expenditures by visits(CD) and the Details of ambulatory care orders(OO) from the LHID2010 will be connected to build up

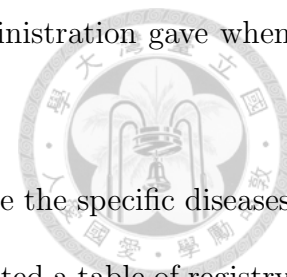
the casual relationship between the drugs that the patients took and the diseases that the patients got after they took the drug.



The CD table mainly records about the information of the patients' treatments and their personal information. The diseases that the patients had been diagnosed are retrieved in this table for the use of understanding what disease the patient might got after taking the drug. Some of the columns are picked from the CD table, including FEE_YM (the year and the month of the fee), APPL_TYPE (apply type), HOSP_ID (the id of the hospital), APPL_DATE (apply date), CASE_TYPE (case type), SEQ_NO (sequence number), FUNC_TYPE (department type), FUNC_DATE (appointment date), ACODE_ICD9_1 (International Classification of Diseases-1), ACODE_ICD9_2 (International Classification of Diseases-2), ACODE_ICD9_3 (International Classification of Diseases-3). The data in ACODE_ICD9_1, ACODE_ICD9_2 and ACODE_ICD9_3 refers to the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) and a domestic code, A-code, which was reorganized from ICD-9 in 1975. (National Institutes of Health, 2014)

The OO table mainly records about the information of the prescription that the doctors prescribed for the patients. The drugs that the doctors prescribed are obtained in this table to find out what the drugs the patient might took that results in the diseases they got afterwards. There are some columns that are picked from the OO table, which includes FEE_YM (the year and the month of the fee), APPL_TYPE (apply type), HOSP_ID (the id of the hospital), APPL_DATE (apply date), CASE_TYPE (case type), SEQ_NO (sequence number), DRUG_NO (the drug number), DRUG_USE (the dosage of the drug), DRUG_FRE (the frequency of the drug taken). The data in DRUG_NO refers

to the identification number that National Health Insurance Administration gave when a new drug is launched. (National Institutes of Health, 2014)



For focusing on serious adverse drug reactions, we need to define the specific diseases that belong to the six categories. The National Health Insurance listed a table of registry for catastrophic illness patients.¹ There are 263 kinds of disease, such as chronic renal failure and cirrhosis of liver. With this table, we can shrink into a smaller database that puts the main issue on serious ADRs.

While there may be lots of different identification number that has similar ingredients and strength, we make use of the data from National Health Insurance Administration to connect the identification number to its corresponding ATC-code from the Anatomical Therapeutic Chemical classification system.² It may help us accurately find the relationship between the drugs and the unexpected serious diseases.

3.3 Data Preparation

In this study, we only make use of the data from January 1st, 1998 through December 31st, 2003. We found that the data before 1998 is incomplete, so we choose to start from 1998. The data between 1999 through 2003 will be the main part to conduct this experiment, and data in 1998 will be an auxiliary variable for observing the reference time T' mentioned in Section 3.1. As for the data after 2003, we discovered that the patients who has catastrophic illness dramatically decreased from 24917 to 2344 patients.

¹https://www.nhi.gov.tw/Content_List.aspx,?n=3AE7F036072F88AF&topn=D39E2B72B0BDF15

²https://www.nhi.gov.tw/Content_List.aspx?n=238507DCFE832EAE&topn=3FC7D09599D25979

In 2004, the National Health Insurance Administration released a program named “Hospital Excellency Project”. This program aims to control the annual medical expenses of individual hospitals. However, some hospitals refused to let catastrophic illness patients see the doctors for their benefits(Taiwan Medical Association, 2005). For the integrity of data, we decided to take the data part between 1998 to 2003.

We connect the CD table and OO table together to be able to see the correlation of the cause-and-effect between the drugs and the disease. Integrating the two tables into a big table may provide patients’ information including both of the data in the CD table and the OO table per year. We linked the two tables by the corresponding columns, which are FEE_YM, APPL_TYPE, HOSP_ID, APPL_DATE, CASE_TYPE, SEQ_NO. We can see that more and more people got disease during 1999 to 2003 in Figure 3.6.

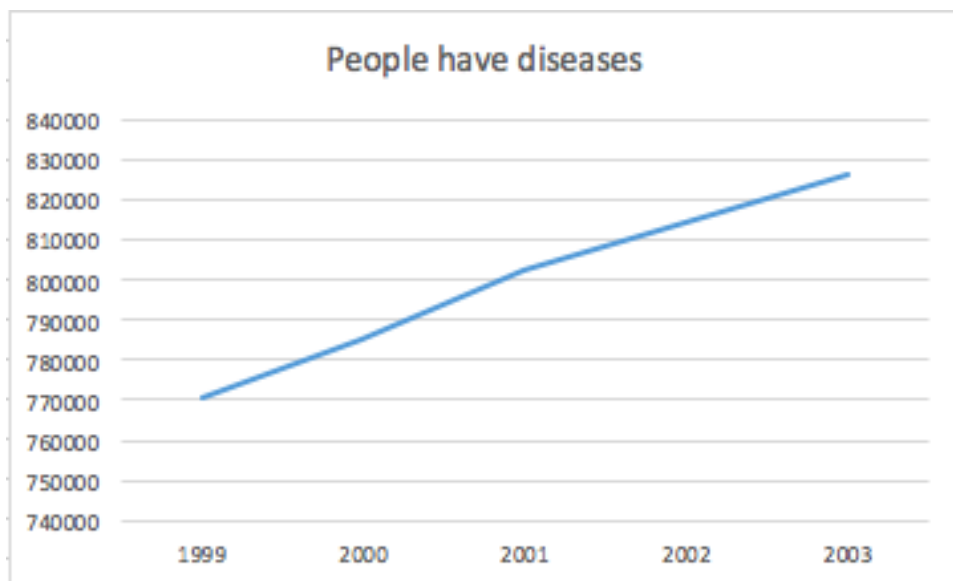


Figure 3.6: People who got disease between 1999 to 2003

As mentioned previously, the CD table has information of the disease the patients got. We established the relationship between a drug and a disease in a record. It will be

more specific to investigate the cause-and-effect of a drug and a disease. There are total 216,899 people who got serious disease during 1999 to 2003 and we have got 19,668 drug identification number (corresponding to 1420 kinds of ATC-code number). We extract the records by these potential drugs and have got 350,035,533 sub-events for the method of TAR. On the other hand, after processing the data set to fit in the method of UTAR, we have got 226,865,574 unexpected sub-events. Figure 3.7 shows the distribution of population who got serious disease in these five years.

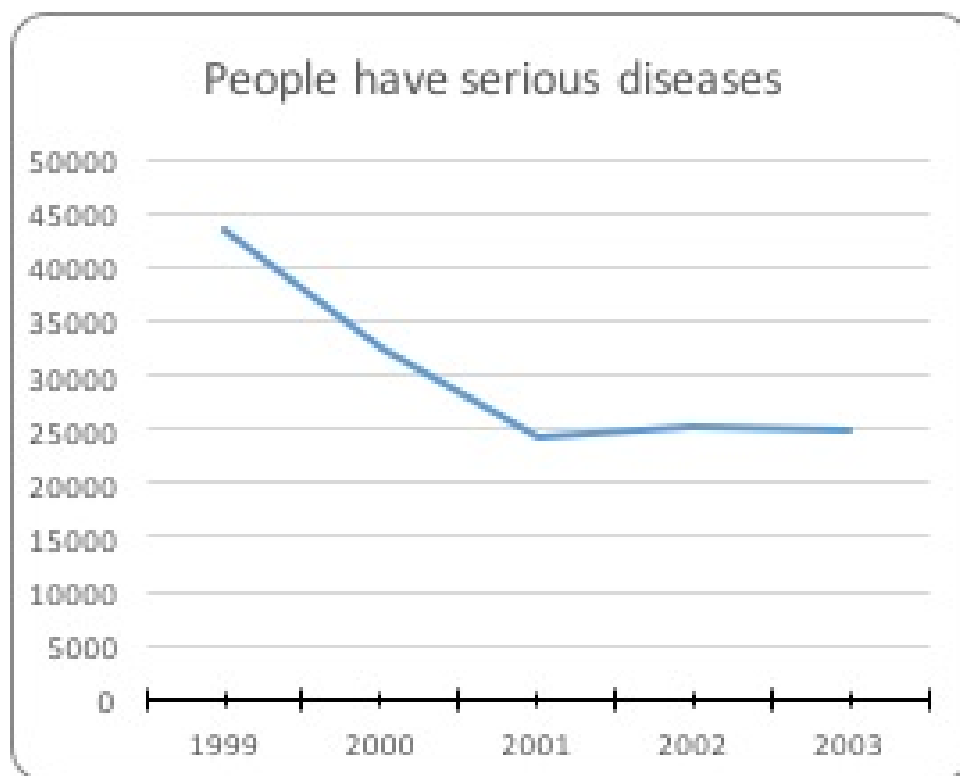
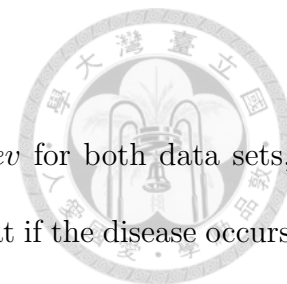


Figure 3.7: People who got serious disease between 1999 to 2003

3.4 Data Analysis



After we calculate the *support*, *confidence*, *leverage*, and *unexlev* for both data sets, the measure *rankRatio* is adopted in this study. We may notice that if the disease occurs before and after the drug, the *unexlev* will decrease compare to its *leverage* and the rank of *unexlev* will be lower than others. Alternatively, if the disease appears after the drug only, it means that the drug has a high probability leading to the disease. In this case, the *unexlev* will not have a significant difference compare to its *leverage* and the rank of *unexlev* will relatively be higher. Therefore, the measurement *rankRatio* is denoted by

$$RR(A \xrightarrow{T} C) = \frac{rank_{leverage}(A \xrightarrow{T} C)}{rank_{unexlev}(A \xrightarrow{T} C)}, \quad (3.7)$$

while $rank_{leverage}(A \xrightarrow{T} C)$ represents the ranks of rule by *leverage* and $rank_{unexlev}(A \xrightarrow{T} C)$ stands for the ranks of rule by *unexlev*.

This ratio may help us rank the potential drugs and its unexpected serious ADRs. If the rank of *unexlev* is high, the *RR* value will be high. This means that this ratio may raise the unexpected rules to a higher rank. Rules that are may be irrelevant will have a lower rank in *RR*.



Chapter 4

Experimental Results

From the experiments above, we've totally got 3,930,750 rules, while 67,173 rules are related to serious diseases. We used the ratio of rank of *leverage* and the rank of *unexlev* to sort the rank of the rules. As mentioned in Section 3.1, there may be positive and negative values of *leverage* and *unexlev*. We selected the rules that both of its values are positive for current study. We found that these selected rules have a positive correlation in both methods. There were total 1,152,200 rules selected, while 21,399 rules are related to serious diseases. For validating our method, we will discuss this topic in the first part. The overall result after implementing this experiment will be discussed in the next section. Since it is the unexpected serious ADRs what we really concern, we will focus on the rules that contain serious diseases in the last part.

Drug Name	ICD-9-CM
A03FA02	794.31
	794.3
	794
M03BX02	794.31
	794.3
	794



Table 4.1: Corresponding names of cisapride, terfenadine, and QT interval prolongation

4.1 Validation

As mentioned in Section 2.2, we want to find out whether cisapride or terfenadine may cause QT interval prolongation. The corresponding drug name and disease name are listed in Table 4.1. Results are shown in Table 4.2. In the case of cisapride, we did find a rule that causes abnormal electrocardiogram (ICD-9-CM: 794.31 , QT interval prolongation belongs to this disease code). However, this rule is not selected since the *leverage* and *unexlev* values are both negative. To find the rules that might have similar meanings, we used a larger category, non-specific abnormal results of function study of cardiovascular system (ICD-9-CM: 794.3). By this way, we found a rule which is ranked 758,877 (approximately 65 percentile of all the selected rules). Yet, we didn't find any rules between cisapride and non-specific abnormal results of function studies (ICD-9-CM: 794). On the other hand, terfenadine was found a rule associated with non-specific abnormal results of function studies, which is ranked 235,043 (approximately 20 percentile of all the selected rules).

For focusing on more serious ADRs, we found a list of drugs published in 2011 that are frequently involved in serious ADRs by a working group associated with the Danish

RR	Rank based on		ATC-code	Drug Name	ICD-9	Disease Name	Unexlev	UTAR support (*N)	UTAR confidence	Leverage	TAR support (*N)	TAR confidence
	Unexlev	Leverage										
758877	1733364	862298	A03FA02	CISAPRIDE	794.3	non-specific abnormal results of function study of cardiovascular system	1.31E-09	1	1.10E-06	2.04E-09	1	7.13E-07
235043	862298	1733364	M03BX02	TERFENADINE	794	non-specific abnormal results of function studies	2.76E-09	1	3.25E-05	2.80E-09	1	2.23844E-05

Table 4.2: Validated rules of cisapride and terfenadine in the selected rules (N = 350,035,533)



Medicines Agency's network "Prevention of Medication Error"¹. There are 20 drugs listed with ATC code and their corresponding serious ADRs. Due to some of the corresponding diseases may adapt to many different categories of ICD-9-CM codes (e.g. the disease with blood clot may match to codes of obstetrical blood-clot embolism or other nonspecific findings on examination of blood), we eliminate them in this study since we are not sure which disease the drug will cause. We selected 13 drugs with its corresponding serious ADRs in Table 4.3. A drug has at most three ICD-9-CM codes for these diseases belong to the same category.

The results showed that there are totally eight rules selected in Table 4.4. The first one is ranked 7573, which is approximately 0.6 percentile by *RR* of all the selected rules. All of the eight rules are related to *cardiac dysrhythmias*, which belongs to ICD-9-CM: 427. These diseases may cause live-saving treatment or prolonged hospitalization. From this validation, we conclude that the method used in NHIRD is significant for finding serious ADRs.

4.2 Overall Results

We listed the top 15 rules ranked by *RR* in Table 4.5. We found that there is an indirect relationship between the drugs and diseases. For example, prescribed drugs that contain *tetanus toxoid* has a relationship with disease that belongs to *open wound* category.

¹https://laegemiddelstyrelsen.dk/en/publications/2011/publication-on-medicines-most-frequently-involved-in-serious-adverse-drug-events/~/_media/D351DCAA2DB4463498724643F4E876C6.ashx



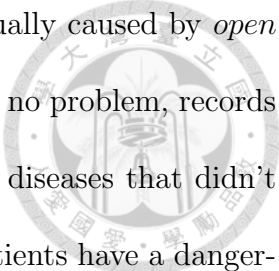
Drug Name	ICD-9-CM
C01BD01	427.89
	427.8
	427
L01XA02	995.91
	995.9
	995
C01AA05	427
J01C	571
C01CA24	427.5
	427
B05BB02	586
N03AB05	427.5
	427
C01DA02	427.5
	427
B01AB10	415.1
	415
C01BB01	427.89
	427.8
	427
C01CA03	427.5
	427
C07AB02	427.89
	427.8
	427
C08CA05	458

Table 4.3: List of drugs that involve serious ADRs with corresponding ICD-9-CM code



RR	Rank based on		ATC-code	Drug Name	ICD-9	Disease Name	Unexlev	UTAR support (*N)	UTAR confidence	Leverage	TAR support (*N)	TAR confidence
	Unexlev	Leverage										
7573	4754	26973	C01AA05	DIGOXIN	427	Paroxysmal supraventricular tachycardia	4.20E-07	121	3.55E-04	7.16E-07	282	3.26E-03
34423	3952	20714	C01DA02	GLYCERYL NITRATE	427	Paroxysmal supraventricular tachycardia	4.54E-07	23	3.02E-04	9.63E-07	403	2.22E-03
40743	11636	59467	C01BD01	AMIODARONE	427.89	Bradycardia	1.42E-07	56	4.55E-03	2.75E-07	100	5.27E-03
47229	8826	44175	C01BD01	AMIODARONE	427	Paroxysmal supraventricular tachycardia	1.93E-07	79	6.41E-03	4.00E-07	147	7.75E-03
57047	61136	297205	C01BD01	AMIODARONE	427.8	Other specified cardiac dysrhythmias	1.77E-08	7	5.68E-04	2.72E-08	10	5.27E-04
92223	196330	865920	C01DA02	GLYCERYL NITRATE	427.5	Cardiac arrest	4.78E-09	2	1.67E-05	5.22E-09	2	1.10E-05
257609	29550	100162	C07AB02	METOPROLOL	427	Paroxysmal supraventricular tachycardia	4.57E-08	2	1.11E-03	1.36E-07	58	2.05E-03
614155	125152	282781	C01BB01	LIDOCAINE	427.89	Bradycardia	7.37E-09	12	5.78E-04	2.94E-08	16	5.55E-04

Table 4.4: Validated rules appear in the selected rules ($N = 350,035,533$)



Tetanus toxoid is a drug that used to prevent tetanus, which is usually caused by *open wound*. While we are confirmed that the pre-process of the data has no problem, records should come from every patient's first taken drugs and unexpected diseases that didn't happen before taking the drug. Therefore, we suspect that these patients have a dangerous working environment, such as construction field. This factor might led to the patients often get *open wound* disease in different parts of body. As for the rule of the third ranking, we observe that the drug of *progesterone* may cause *antepartum examination*. We think that some of the women may want to do In Vitro Fertilization (IVF), so they took the *progesterone* for stable female hormones and then have *antepartum examination* among the next six months. As we know that the *progesterone* would not lead to *antepartum examination*, there is still a phenomenon that *antepartum examination* happens after taking *progesterone* for those patients. These 15 rules are just the top 0.001% of all of the selected rules. However, we want to focus on the ranks of serious ADRs.

4.3 Description of Potential Rules with Serious Diseases

As mentioned above, we've obtain 21,399 rules that are related with serious diseases. In these rules, there are 1371 kinds of drug and 231 kinds of serious diseases. The related records are fetched if they satisfy these rules. We like to see the description of these records.

In the beginning, the amount of records and patients in each year is concerned. Figure

RR	Rank based on		ATC-code	Drug Name	ICD-9	Disease Name	Unexlev	UTAR support (*N)	UTAR confidence	Leverage	TAR support (*N)	TAR confidence
	Unexlev	Leverage										
1	3	343	J07AM01	TETANUS TOXOID	883.0	Open wound of finger without complication	2.32E-05	9670	8.06E-03	2.61E-05	9973	8.31E-03
2	6	599	J07AM01	TETANUS TOXOID	884.0	Open wound of upper extremity (upper limb)	1.67E-05	7343	6.12E-03	1.94E-05	7609	6.34E-03
3	7	524	G03DA04	PROGESTERONE	V22.1	Antepartum examination (Supervision of other normal pregnancy)	1.66E-05	6549	1.80E-02	2.06E-05	7648	2.10E-02
4	5	361	J07AM01	TETANUS TOXOID	A501	Open wound of upper extremity (upper limb)	2.15E-05	9146	7.62E-03	2.54E-05	9821	8.18E-03
5	8	538	J07AM01	TETANUS TOXOID	A502	Open wound of lower extremity (lower limb)	1.63E-05	7427	6.19E-03	2.03E-05	8082	6.73E-03
6	12	731	J07AM01	TETANUS TOXOID	891.0	Open wound of knee, leg and ankle without complication	1.48E-05	6594	5.49E-03	1.73E-05	6807	5.67E-03
7	4	242	S01FA06	TROPICAMIDE	367.1	Myopia	2.20E-05	8608	2.37E-02	3.19E-05	11756	3.24E-02
8	11	640	J07AM01	TETANUS TOXOID	894.0	Open wound of lower extremity (lower limb)	1.53E-05	7175	5.98E-03	1.86E-05	7491	6.24E-03
9	19	892	J07AM01	TETANUS TOXOID	873.40	Open wound of face without complication	1.30E-05	5781	4.82E-03	1.51E-05	5941	4.95E-03
10	21	985	J07AM01	TETANUS TOXOID	882.0	Open wound of hand without complication	1.24E-05	5257	4.38E-03	1.41E-05	5439	4.53E-03
11	23	966	J07AM01	TETANUS TOXOID	A500	Open wound of face	1.20E-05	5269	4.39E-03	1.43E-05	5594	4.66E-03
12	31	1247	J07AM01	TETANUS TOXOID	873.0	Open wound of face without complication	1.05E-05	4602	3.83E-03	1.21E-05	4735	3.95E-03
13	16	619	S01FA06	TROPICAMIDE	367.9	Disorder of refraction and accommodation, Refractive error	1.36E-05	5447	1.50E-02	1.90E-05	7053	1.94E-02
14	9	309	G03DC02	NORETHINDRONE ACETATE	626.4	Irregular menstrual cycle	1.61E-05	7764	1.74E-02	2.74E-05	10865	2.43E-02
15	2	64	J01AA02	DOXYCYCLINE (HCL)	706.1	Acne vulgaris	2.41E-05	14159	1.20E-02	5.56E-05	23050	1.95E-02

Table 4.5: Rules of top 15 base on RR (N = 350,035,533)

4.1 shows that the amount of records increased dramatically through July 1st, 1998 to June 30th,2000. Yet, the records remain nearly the same in the years after. Figure 4.2

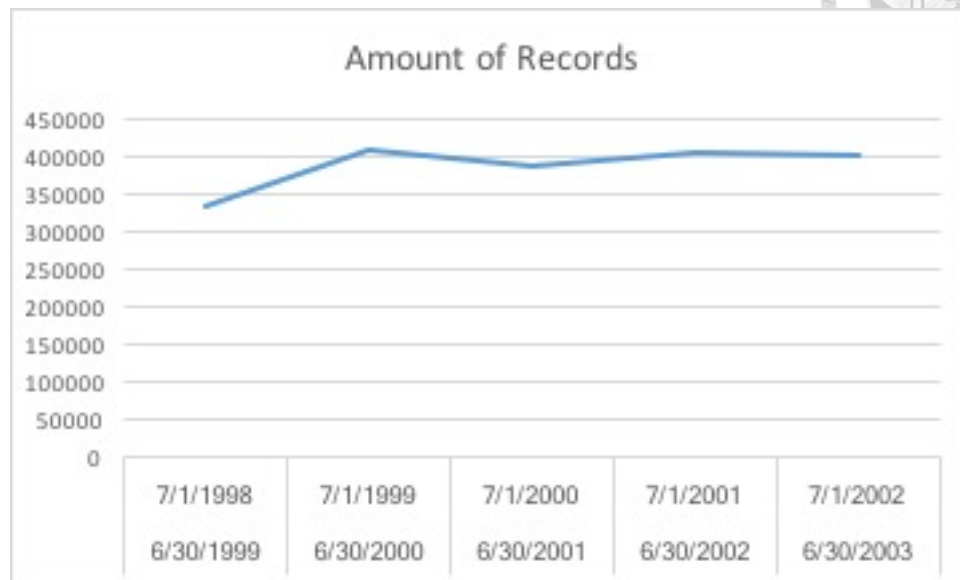


Figure 4.1: Amount of Records in each year

displays patients are continuing increased during the years.

The distribution of sex type in these related records is shown in Figure 4.3. The proportion of female and male are nearly the same. We want to know whether the distribution of the sex type in these records remain the same. So, we conduct a Z test with

$$H_0 : p_0 = p_1$$

$$H_1 : p_0 \neq p_1,$$

where p_0 represents the proportion of female in the population and p_1 is the proportion of female in the related records. The $p - value$ is 0.53, so we don't reject H_0 . Unlikely, there is no significant difference between the records and the population. Table 4.6 shows the amount and proportion of sex type.

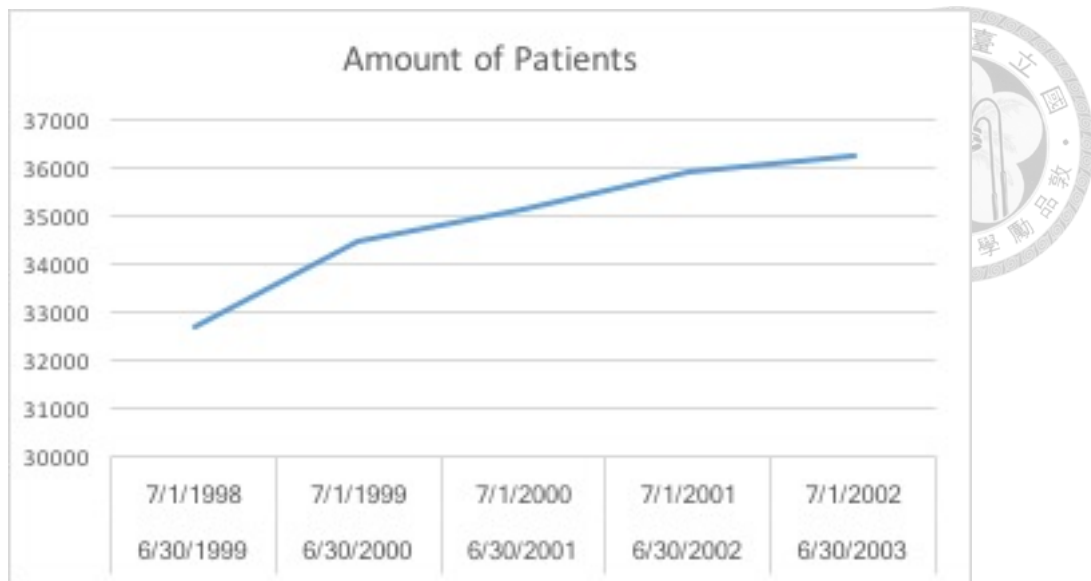


Figure 4.2: Amount of Patients in each year

sex type	sample	%	population	%
Female	20606	50.66	447813	50.63
Male	20071	49.34	436681	49.37

Table 4.6: The proportion of sex type

In these related records, Figure 4.4 shows the distribution of age. We grouped the age into five groups, defined in Table 4.7. The majority are adults in these records. We also want to know whether the distribution of age in the records remain the same. The chi-square goodness-of-fit test is applied with

H_0 : The distributions remain the same

H_1 : The distributions are not the same.

The $p - value$ is 0.34, so we don't reject H_0 as well. Unfortunately, the distributions of the records and the population remain the same. Table 4.8 shows the amount of people in each age group.

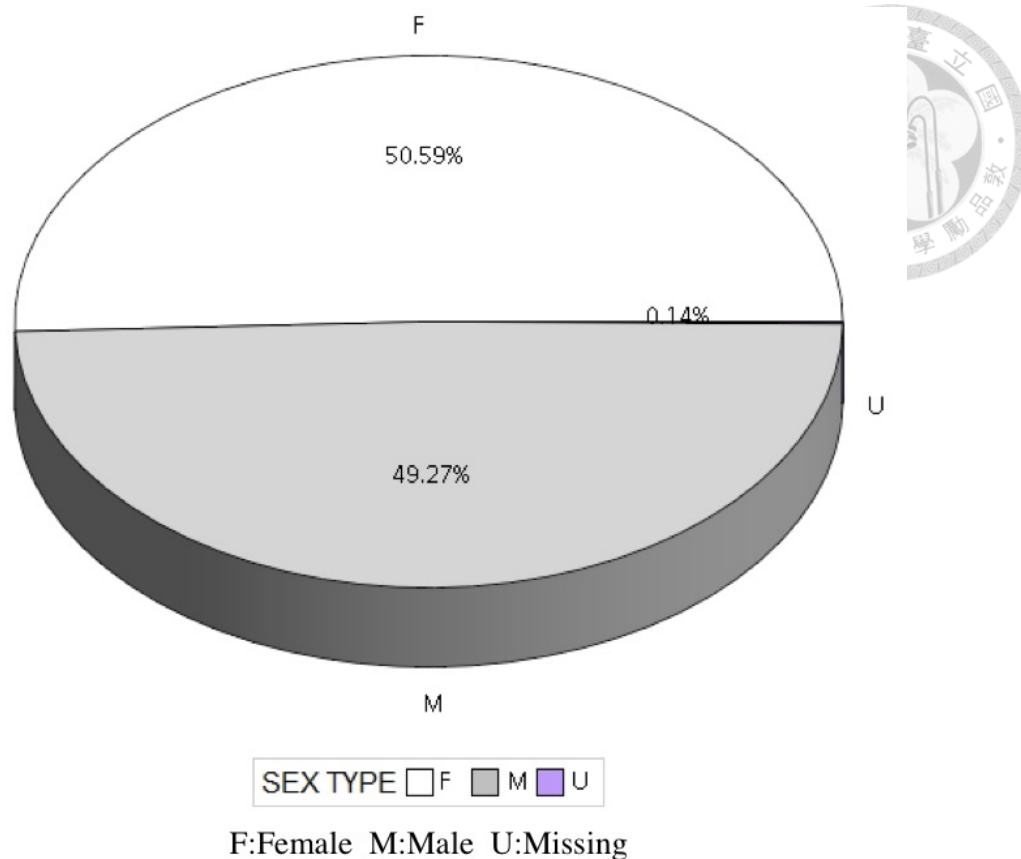


Figure 4.3: Distribution of Sex Type

Drugs are also an interesting variable we want to observe. As defined by World Health Organization Collaborating Centre (WHOCC), the ATC system categorized the drugs into 14 main groups.² The categorized groups are shown in Table 4.9. The distribution of taken drugs in the records are shown in Figure 4.5. We may notice that drugs that belongs to “A” and “R” are taken the most. We make use of these groups to discover whether the distribution of taken drugs remains the same. The chi-square goodness-of-fit test is conducted with

$$H_0 : \text{The distributions remain the same}$$

²https://www.whooc.no/filearchive/publications/2017_guidelines_web.pdf

Age Range	Group Name
0-17	child
18-65	adult
66-79	senior
80-99	old
100+	centenarian



Table 4.7: Age Group

	child	adult	senior	old	centenarian	sum
records	2021	33020	5039	2928	211	43219
population	43509	720495	108242	62511	4823	939580

Table 4.8: The amount of people in records and population

H_1 : The distributions are not the same.

Table 4.10 shows the amount of records taking drugs. The $p - value$ is smaller than $2.2E-16$, so we reject H_0 . We've found that the proportion of Group "C" increased in the records. This caused our interest to understand the related informations of Group "C".

There are 3,224 rules that are related with cardiovascular-system-drugs and serious diseases. Among them, there are 156 kinds of drugs and 200 kinds of serious diseases. Similar to the way we mentioned above, we further selected the records that belongs to Group "C". In Figure 4.6 and Figure 4.7, we found that the amount during July 1st, 2002 to June 30th, 2003 increased in a sudden. We think that the implement of using ICD-9-CM code instead of A-code enforced by National Health Insurance Administration is the main reason. Figure 4.8 shows that female are more than male in the case of



Represented Code	Anatomical Main Group
A	Alimentary tract and metabolism
B	Blood and blood forming organs
C	Cardiovascular system
D	Dermatologicals
G	Genitourinary system and sex hormones
H	Systemic hormonal preparations, excl. sex hormones and insulins
J	Antiinfectives for systemic use
L	Antineoplastic and immunomodulating agents
M	Musculo-skeletal system
N	Nervous system
P	Antiparasitic products, insecticides and repellents
R	Respiratory system
S	Sensory organs
V	Various

Table 4.9: ATC-code Classification Principle



	records	population
A	1494371	32398092
B	216764	4694976
C	196344	4155700
D	178283	3931139
G	108579	2378852
H	112219	2433559
J	615237	13437535
L	2071	42762
M	657468	14219353
N	688278	14916697
P	11589	259711
R	1689294	36858896
S	208103	4485639
V	2168	45405
sum	6180768	134258316

Table 4.10: Amount of records taking drugs classified by ATC Group

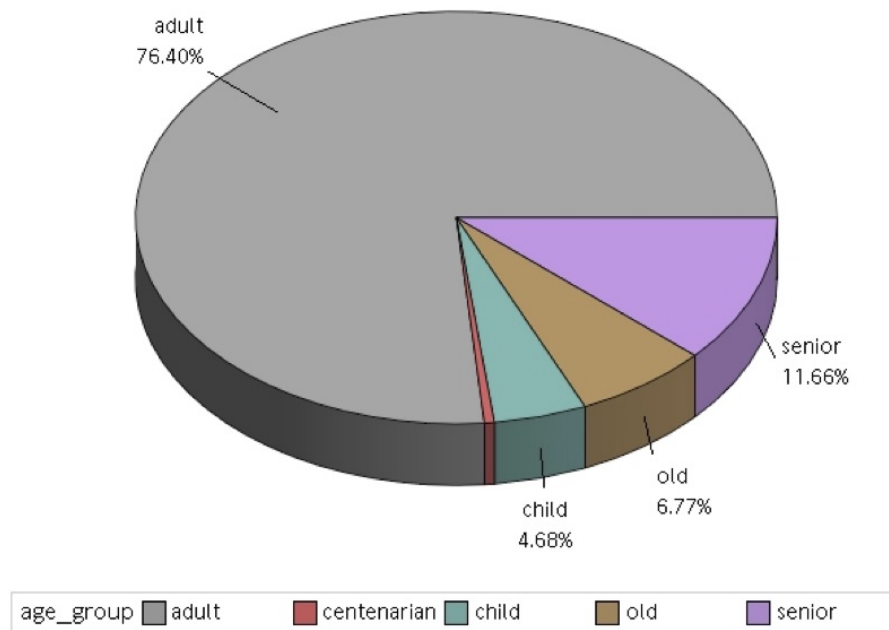


Figure 4.4: Distribution of Age

taking drugs in Group “C”. In Figure 4.9, we may notice that although adults are still the majority, the proportion of adult decreased while the proportion of senior increased. This might inform us that the much more senior happens to take drugs for cardiovascular system.

4.4 Results of Potential Rules with Serious Diseases

In this section, we first listed the top 15 rules ranked by *RR* that is related to serious diseases. We found that most of the drugs are used to cure certain diseases, but there are other diseases that might have relationships with these certain diseases. Take the rule of third ranking in Table 4.11 as an example, we found that drugs that contain *gemfibrozil* has a relationship with *pure hypertriglyceridemia*. *Gemfibrozil* is used to treat high

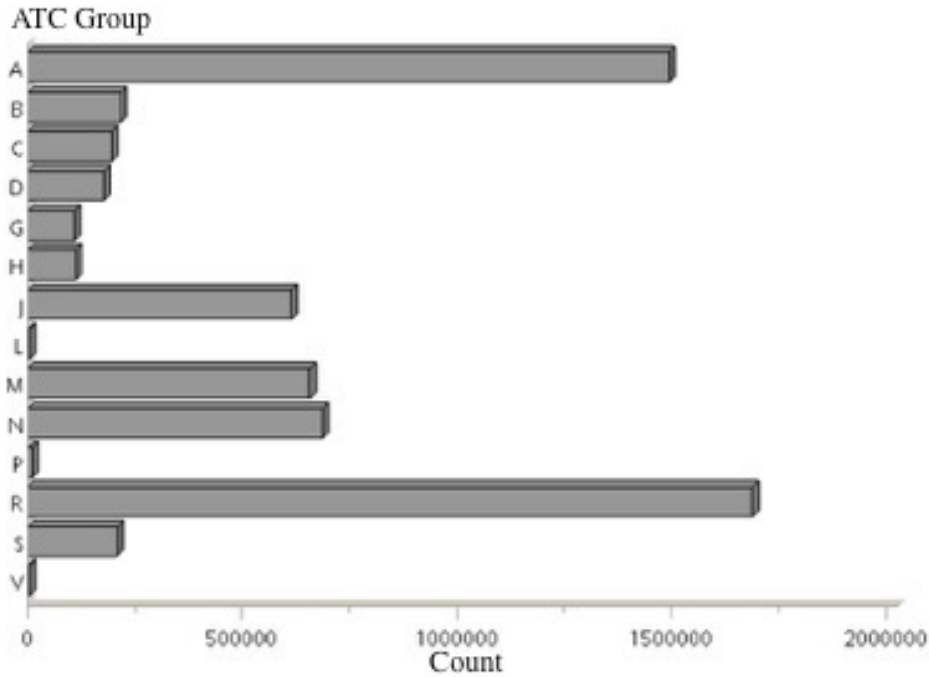


Figure 4.5: Distribution of taken drugs

level of cholesterol and triglyceride in the blood, and *pure hypertriglyceridemia* is a disease that happens when triglyceride levels are elevated. When a patient took the drug for curing high level of cholesterol and triglyceride for a period of time, the amount of cholesterol in the blood lowered down but the triglyceride still remained high. Thereby, the doctor may diagnose *pure hypertriglyceridemia* in the next record. Another example is the rule with rank number four. This drug is composed of three ingredients, including *glycerin*, *sodium chloride*, and *fructose*. It is used to reduce intracranial pressure or to cure the cerebral edema. Usually, patients with a high risk of cerebrovascular event are prescribed for this drug. When the patients happen any event such as not taking the drug on time, they might get *cerebrovascular accident*. However, the patterns of these examples are not what we intended to study. Similar to Section 4.2, these events can be explained as the diseases might happen after taking the drug. Yet, what we want is the

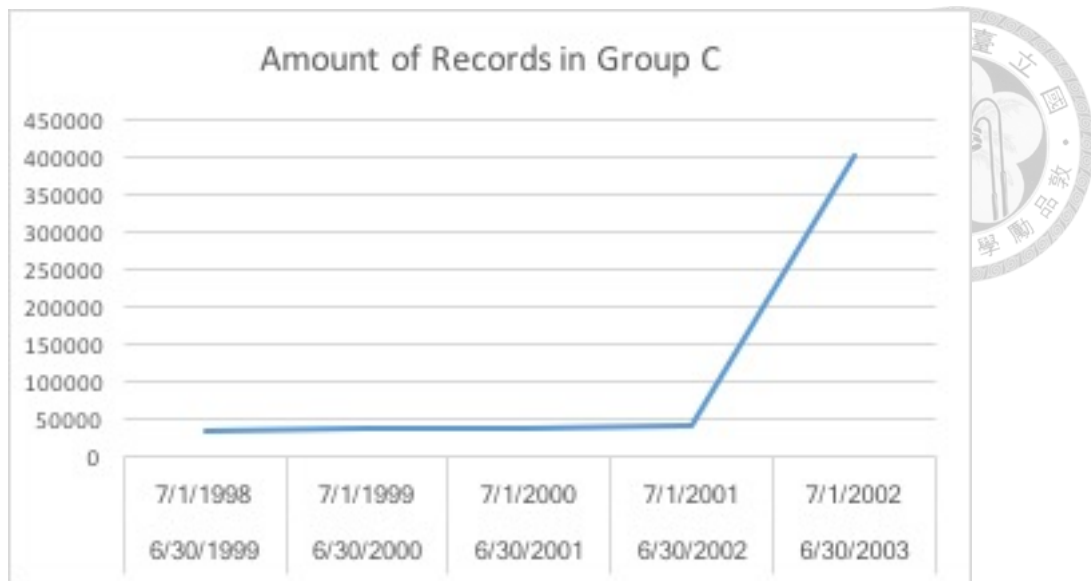


Figure 4.6: Amount of Records of Group “C” in each year

pattern that causes serious ADRs. These 15 rules are just the top 0.07% of the selected rules that cause severe diseases.

RR	Rank based on		ATC-code	Drug Name	ICD-9	Disease Name	Unexlev	UTAR support (*N)	UTAR confidence	Leverage	TAR support (*N)	TAR confidence
	Unexlev	Leverage										
382	2985	30412	B05BC01	MANNITOL	431	Intracerebral hemorrhage; Cerebral hemorrhage	5.94E-07	212	1.00E-02	6.27E-07	222	1.05E-02
501	3923	37421	B05BC92	GLYCERIN (=GLYCEROL), SODIUM CHLORIDE, FRUCTOSE (=LAEVULOSE)	431	Intracerebral hemorrhage; Cerebral hemorrhage	4.57E-07	162	1.53E-02	4.90E-07	173	1.64E-02
531	808	7625	C10AB04	GEMFIBROZIL	272.1	Pure hypertriglyceridemia	1.67E-06	656	4.11E-03	2.65E-06	973	6.10E-03
1309	11172	90358	B05BC92	GLYCERIN (=GLYCEROL), SODIUM CHLORIDE, FRUCTOSE (=LAEVULOSE)	436	Cerebrovascular accident; Acute cerebral vascular disease	1.49E-07	57	5.39E-03	1.57E-07	58	5.49E-03
1322	7462	60242	D02AE01	UREA	757.39	Other specified anomalies of skin	2.32E-07	91	7.22E-04	2.70E-07	100	7.93E-04
1416	6440	51417	D10AD01	RETINOIC ACID (=TRETINOIN)	757.39	Other specified anomalies of skin	2.72E-07	104	9.03E-04	3.31E-07	121	1.05E-03
1565	10596	83398	D01AE12	SALICYLIC ACID	757.39	Other specified anomalies of skin	1.58E-07	60	9.99E-04	1.75E-07	64	1.07E-03
1586	5320	41796	G04CA03	TERAZOSIN (HCL 2H2O)	185	Cancer of prostate	3.33E-07	128	7.97E-04	4.29E-07	158	9.84E-04
1685	16726	130230	B05BC01	MANNITOL	430	Subarachnoid hemorrhage	9.24E-08	34	1.60E-03	9.44E-08	34	1.60E-03
2060	18222	137689	B05BC92	GLYCERIN (=GLYCEROL), SODIUM CHLORIDE, FRUCTOSE (=LAEVULOSE)	430	Subarachnoid hemorrhage	8.34E-08	30	2.84E-03	8.72E-08	31	2.93E-03
2130	13733	103252	D02AE01	UREA	757.1	Ichthyosis congenita	1.17E-07	43	3.41E-04	1.31E-07	47	3.73E-04
2310	18479	137300	R06AE51	BUCLIZINE 2HCL, NIACIN (=NICOTINIC ACID)	277.9	Unspecified disorder of metabolism	8.19E-08	31	1.05E-03	8.75E-08	32	1.08E-03
2336	8436	62574	H03AA01	LEVOTHYROXINE SODIUM	243	Congenital hypothyroidism	2.03E-07	72	1.42E-03	2.58E-07	91	1.80E-03
3126	18869	134296	B02BA01	PHYTOMENADIONE (=VIT K1)	431	Intracerebral hemorrhage; Cerebral hemorrhage	7.99E-08	33	1.23E-03	9.03E-08	35	1.30E-03
3296	1558	11006	B05XA03	SODIUM CHLORIDE	431	Intracerebral hemorrhage; Cerebral hemorrhage	1.05E-06	619	4.58E-04	1.86E-06	824	6.09E-04

Table 4.11: Rules of top 15 base on RR in serious diseases (N = 350,035,533)

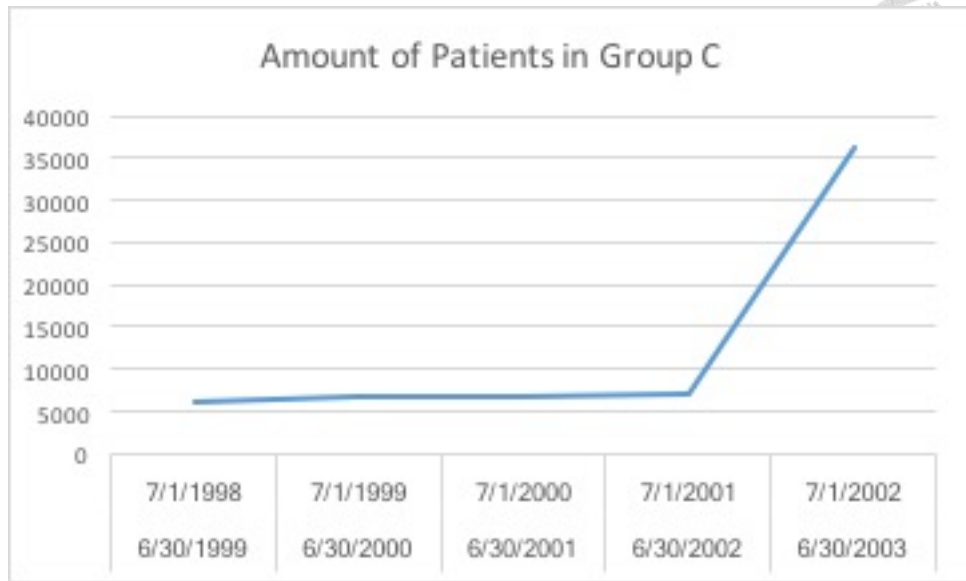


Figure 4.7: Amount of Patients of Group “C” in each year

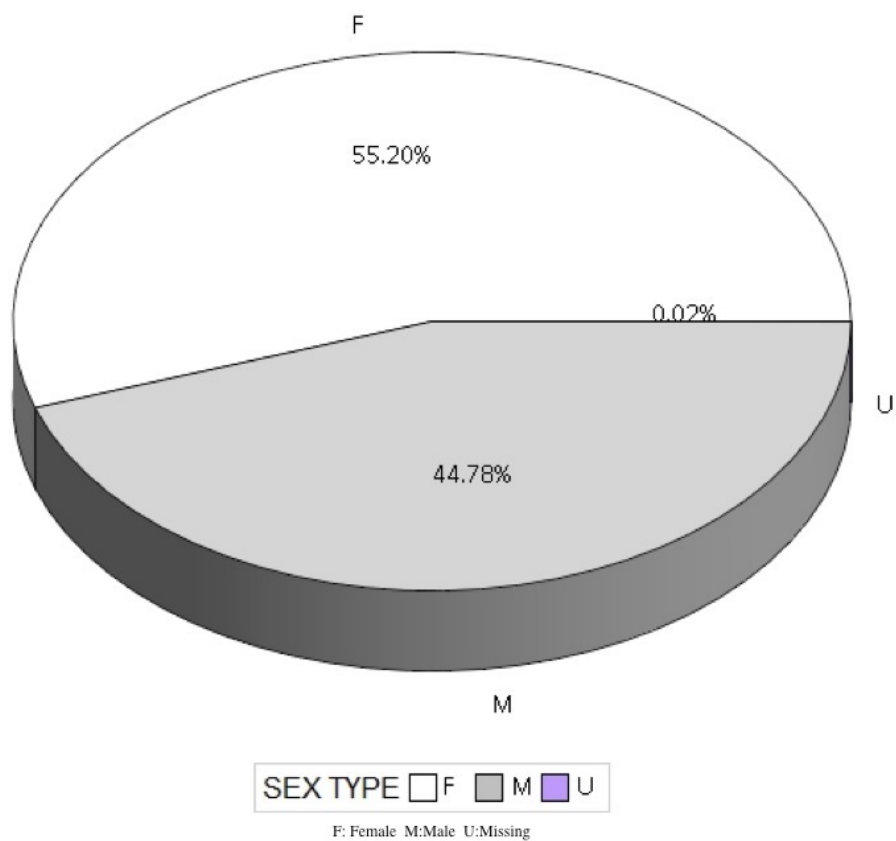


Figure 4.8: Distribution of Sex Type in Group “C”

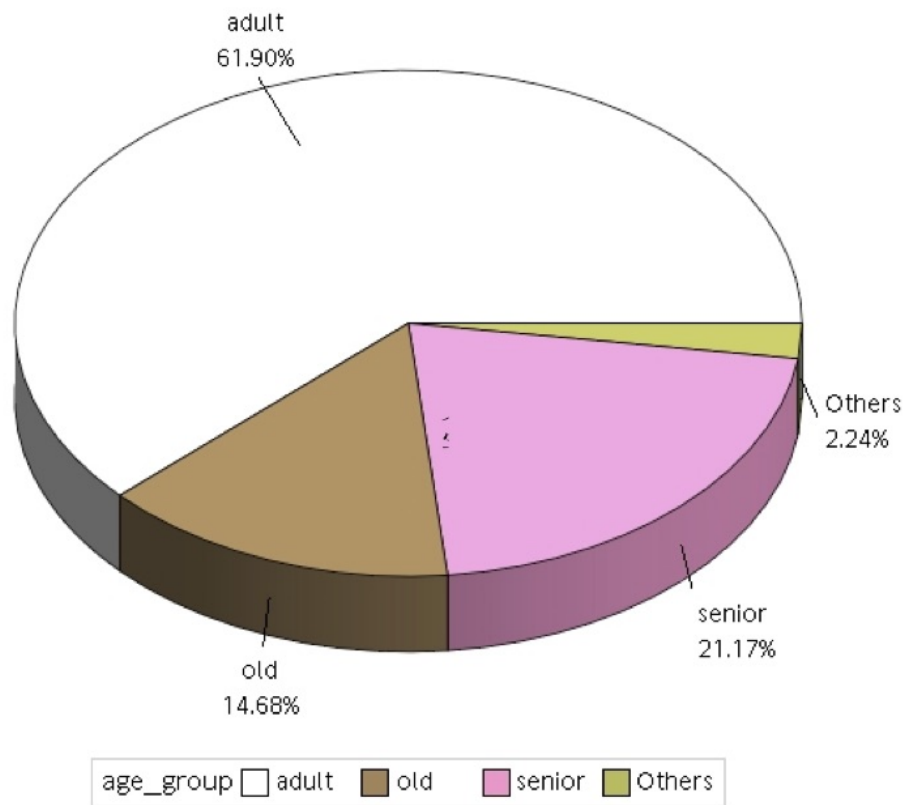


Figure 4.9: Distribution of Age Group in Group “C”



Chapter 5

Conclusion

5.1 Contribution

Nowadays, more and more people suffered from serious ADRs. There are many methods, such as the reporting system used to detect serious ADRs. However, the traditional ways might be inefficient. In this study, we used the method proposed by McAullay et al. (2009) to find unexpected, serious ADRs from NHIRD. We started from the initializing step to the last step, UTAR. This method is an effective way to find unexpected, serious ADRs as shown in Chapter 4. The *RR* ranking system help us detect the unexpected rules. By making use of the *leverage* and *unexlev*, we could eliminate 2,778,550 rules, and select the 1,152,200 rules (about 30% of among all the rules) for detecting ADRs. While focusing on serious ADRs only, we further pull out the rules that are related with serious diseases. Even though there might be patterns we are not interested, they are just the tip of an iceberg. There are 26 rules listed in Table 4.3, while eight of them were found in a high rank. The data we used span from 1998 to 2003. In Table 4.2, we

may notice that rules of *cisapride* and *terfenadine* were selected. These rules may be examined in the early of 2004 after we conduct this experiment. However, these drugs were examined in the middle of 2004 and 2005 and were removed respectively. We believe that unexpected, serious ADRs will be detected more accurately and efficiently by this method.

5.2 Limitation and Future Work

Even though we found out some serious ADRs in a high rank, this study still has some limitations. The uncertainty of patients can't be defined. Some patients listen to the doctors and take the drugs on time, while some patients just see the doctors and put the drugs aside. In this database, we won't know whether the patient really take the drug. What we may know is the drugs and diseases that was prescribed and diagnosed for the patient. There is another limitation. For only using NHIRD in this study, we only obtain the rules that are related to Taiwanese. There may be some genetic differences between different races. Thus, we may not acquire all of the serious ADRs in the world.

In this study, we only discuss about the relationships merely between drugs and diseases. We think that this study can be extended by adding features, such as the interactions between the drugs in further research. This study can also be extended by discussing rules generated with different intervals of T and T' . On the basis of this study, we believe that methods of detecting serious ADRs will be more complete in the future. In the medical aspect, these selected rules may be examined by experts. If there are any unexpected, serious ADR found in these rules, the casualties may decrease and may have

a new finding in medical field.





Bibliography

- Bonn, Dorothy. 2005. Adverse drug reactions remain a major cause of death. *The Lancet* **351**(9110) 1183.
- Bresalier, Robert S., Robert S. Sandler, Hui Quan, James A. Bolognese, Bettina Oxenius, Kevin Horgan, Christopher Lines, Robert Riddell, Dion Morton, Angel Lanas, Marvin A. Konstam, John A. Baron. 2005. Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. *New England Journal of Medicine* **352**(11) 1092–1102.
- Chan, Agnes L. F., Haw Yu Lee, Chi-Hou Ho, Thau-Ming Cham, Shun Jin Lin. 2008. Cost evaluation of adverse drug reactions in hospitalized patients in taiwan: A prospective, descriptive, observational study. *Current Therapeutic Research* **69**(2) 118–129.
- Chazard, Beuscart, Emmanuel, Gregoire Ficheur, Stephanie Bernonville, Michel Luyckx, Regis. 2011. Data mining to generate adverse drug events detection rules. *IEEE Transactions on Information Technology in Biomedicine* **15**(6) 8.
- Hsieh, Tsai-Hsuan. 2014. Detecting drug safety signals from national taiwan health insurance research database : A learning to rank approach. Thesis, National Taiwan University.

Katzung, Bertram G. 2015. *Introduction: The Nature of Drugs & Drug Development & Regulation*. McGraw-Hill Medical, New York, NY.



McAullay, Damien, Chris Kelman, Jie Chen, Huidong (Warren) Jin, Christine M. O'Keefe, Hongxing He. 2009. Signaling potential adverse drug reactions from administrative health databases. *IEEE Transactions on Knowledge & Data Engineering* **22**. doi:10.1109/TKDE.2009.212.

Moore, Nicholas, Dominique Lecointre, Catherine Noblet, Michel Mabile. 1998. Frequency and cost of serious adverse drug reactions in a department of general medicine. *British Journal of Clinical Pharmacology* **45**(3) 301–308.

National Health Insurance Administration. 2015. National health insurance in taiwan 2015-2016 annual report. Report, National Health Insurance Administration Ministry of Health and Welfare.

National Institutes of Health. 2014. Nhir codebook. *NHIRD Codebook* **103** 179.

Reps, Jenna Marie, Jonathan M. Garibaldi, Uwe Aickelin, Daniele Soria, Jack Gibson, Richard Hubbard. 2013. Comparison of algorithms that detect drug side effects using electronic healthcare databases. *Soft Computing* **17**(12) 2381–2397.

Shaikh, W. A. 2000. The changing face of antihistamines and cardiac adverse drug reactions: a clinical perspective. *Journal of the Indian Medical Association* **98**(7) 397–399.

Taiwan Medical Association. 2005. The impact and response of the hospitals in hospital excellency project. *Taiwan Medical Journal* **48**(6).

Waller, Patrick. 2009. *Basic Concepts*. Wiley-Blackwell.

Wang, S. H., C. Y. Lin, T. Y. Huang, W. S. Wu, C. C. Chen, S. H. Tsai. 2001. Qt interval effects of cisapride in the clinical setting. *International Journal of Cardiology* **80**(2) 179–183.

