BUILDING AN AUTOMATED ANALYSIS SYSTEM TO ANALYZE CORRELATION BETWEEN TWO DISEASES USING A COHORT STUDY DESIGN AND A LONGITUDINAL HEALTH INSURANCE DATABASE

[1]CHENG-WEI CHANG, [2]LI-CHING WU, [3]TSUNG-WEI LIN, [4]JORNG-TZONGHORNG

[1]Department of Information Management, Hsing Wu University, Taipei, 244, Taiwan,
[2]Department of System Biology and Bioinformatics, National Central University, Jhongli 32001, Taiwan,
[3,4]Department of Computer Science and Information Engineering, National Central University, Jhongli 32001, Taiwan,
[4]Department of Bioinformatics and Medical Engineering, Asia University, Taichung 41354, Taiwan
E-mail:[1]95010@mail.hwu.edu.tw, [2]Ri-chard@mail.sybbi.ncu.edu.tw, [3] loadding000@hotmail.com,
[4]horng@db.csie.ncu.edu.tw

Abstract - In Taiwan, the National Health Insurance Research Database (NHIRD) plays an important role in medical research. NHIRD studies have grown rapidly both in terms of publication number and quality since 2000. The Longitudinal Health Insurance Database (LHID) is a data subset of the NHIRD that was specifically constructed for research purposes. The data forming the LHID is both important and numerous and are many researchers utilizing the NHIRD or the LHID as a research resource now. Among published journals in 2014, the cohort study design is widely used for analyzing data from the NHIRD. With the growth of information in medical databases, efficiency when carrying out research using a medical database is becoming important. However, we have been unable to identify any Taiwanese publications that have used an automated analysis system with either the NHIRD or the LHID A lot of useful information is held in the NHIRD and therefore it is important to discover such information in a speed and efficient manner. In our study, we have implemented an automatic analysis system that is able to analyze correlations between any two diseases using a cohort study design and the LHID.

Index Terms - Cohort Study, Expert System, LHID, NHIRD

I. INTRODUCTION

TAIWAN launched the National Health Insurance (NHI) program on March 1995 and by 2014 almost 99.9% of Taiwan's population is enrolled in this program. The National Health Insurance Research Database (NHIRD) is a product of the NHI program and the NHIRD is the most commonly used and largest medical database in Taiwan. The NHIRD has also been used to create a number of data subsets, one of which is the Longitudinal Health Insurance Database (LHID). This was constructed specifically for research purposes. The LHID contains 1,000,000 beneficiaries of the NHIRD that have been enrolled by random selection from the complete NHIRD Registry of Beneficiaries (ID) [1].

The NHIRD contains important and numberous datasets and therefore many researchers are selecting the NHIRD or the LHID as a research resource. Every year hundreds of new publications that cited the NHIRD as a data source can be identified using PubMed, a service provided by the National Center for Biotechnology Information (NCBI). There were 671 publications identidied by PubMed and citing the NHIRD published in 2014 and of these 376 involved a cohort study design[3]. This findings show that the NHIRD is a highly credible research resource. Cohort studies are a special type of observational study design and this approach is able to improve various types of studies inkling case series, case reports, and cross-sectional analyses; such an approach is very useful and applicable when using a longitudinal database such as the NHIRD [1],[4]. Some variables, such as diseases and drugs are able to use a cohort study design; however, cohort study designs in some cases can be very complicated. [3].As time goes on, it is clear that the NHIRD will gather even more medical information about patients and disease. As a result of this growth in the information stored within the NHIRD, the efficient use of the NHIRD for research purposes is becoming a significant concern.

In spite of the fact that cohort studies are popular and are quite suitable when carring out research using a medical database, a lot of time and labor is still required in order to deal with available complex medical data. Therefore, we need a fast and efficient means of discovering this information. In such circumstances an automatic analysis system ought to be able to reduce the laborious repeated procedures used in such circumstances. This will make research faster and avoid some of the mistakes that sometimes appear when research is published.

In our study, we have implemented an automatic analysis system that can be used to analyze correlations between two diseases using a cohort study design and the LHID. The system developed during our study is expected to accelerate the efficiency of research when it involves the relationship between any two diseases using the medical results held within the National Health Insurance system.

II. MATERIALS

2.1 Cohort Study Design

Cohort studies are a type of observation study and are commonly used in the medical research; a good cohort study design is able to provide powerful results [5]. In our study, the study population used refers to the patients in the NHI medical database.The three fundamental steps needed when carryingout a cohort studies ar shown below [4],[5]:

1. Identified two cohorts that form an exposed group and an unexposed group; these need to be free of outcome at the start of the study.
2. After this grouping, the two cohorts are followed over time to identify the evolution of the outcome and this lasts until the end time point of the database.
3. Finally, compar the risk of outcome between two cohorts.

In cohort studies these three fundamental steps are always included and we have collected some research publications that have focused on diseases, cohort study design and medical databses [6]-[15]; these are summarized in Figure 1. We have followed these procedures when developing our analysis system.

For example, take two diseases, chronic obstructive pulmonary disease (COPD) and lung cancer, as example. First we need to identify two cohorts. One is a group of pateients with COPD, and the other is a group of patietns without COPD. Second, we create two cohorts that fulfill the inclusion criteria. Some of the common inclusion criteria used are sex, age, onset of COPD by the study subject; furthermore, in this case an obvious exclusion criterion is an onset of lung cancer before the onset of COPD. Third, the two cohorts are matched according to sex, age group, and index date. Finally, we follow two cohorts over time and compare the risk of lung cancer between the two cohorts. Some other details, such as confounding factors may need to be taken into account.

In our study, the entire medical data source is retrieved from Longitudinal Health Insurance Database (LHID), which is a subset of National Health Insurance Research Database (NHIRD) and is maintained by the Bureau of National Health Insurance (BNHI) in Taiwan. Systematic sampling is used to create the LHID subset of data and it has been shown that there is no statistically significant differences between the LHID and the NHIRD [1],[6],[7]. The LHID includes the medical records of outpatient visits, hospital admissions, prescriptions, operation codes and diagnosis codes [1],[8]. These records are linkable to each patient through an encrypted personal identification number [9],[10]. Within the LHID, the diagnosis code is according to the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). Some data within the LHID is either missing or incomplete; this can include an unknown gender, the wrong function date or a wrong birth date. When our system carries out an analysis, it excludes records containing incorrect data.

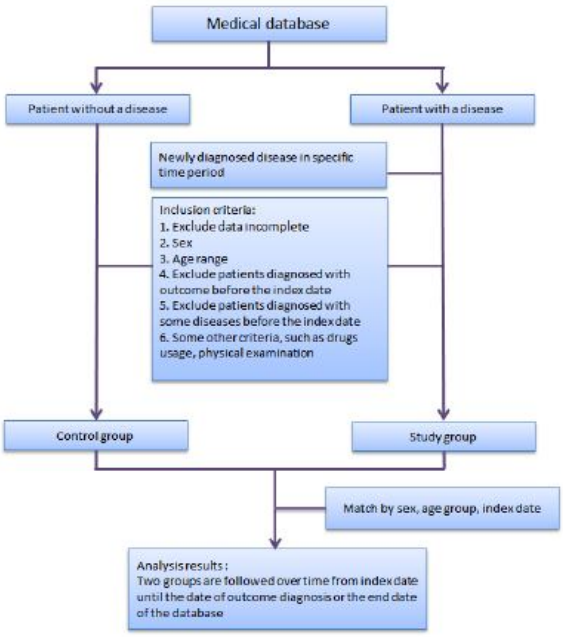


Figure 1. Common procedures used for a cohort study design.

III. Methods

3.1Details of the System Design

The flow of our proposed system is shown in Figure 2. Five components form the system flow; these are website query, data converter, statistical software, results convertor and website displayer.

3.2Website Query

We have classified the parameters into various categories. There are four categories, namely medical records, main disease, outcome of interest and confounding factors. Medical records have several properties, including the start year of the record, the end year of the record, and outpatient/inpatient visits. The main disease is the disease that we intend to observe and this disease is usually believed to develop into outcome of interest. Main disease has a number of properties, for example the age of onset of the illness and the onset year of the disease. The outcome of interest and the confounding factors have similar properties to the main disease such as diagnosis codes and number of diagnosis in a period of time.

Figure 3 outlines the query procedure.Firstly, we needed to choose the years of the data that are to be examined and secondly, we needed to select appropriate information on the main disease, namely the name of the main disease and its diagnosis code. After each step is performed, we check the integrity of the query to avoid there being any unreasonable input. Once complete we send the query on to the next part.

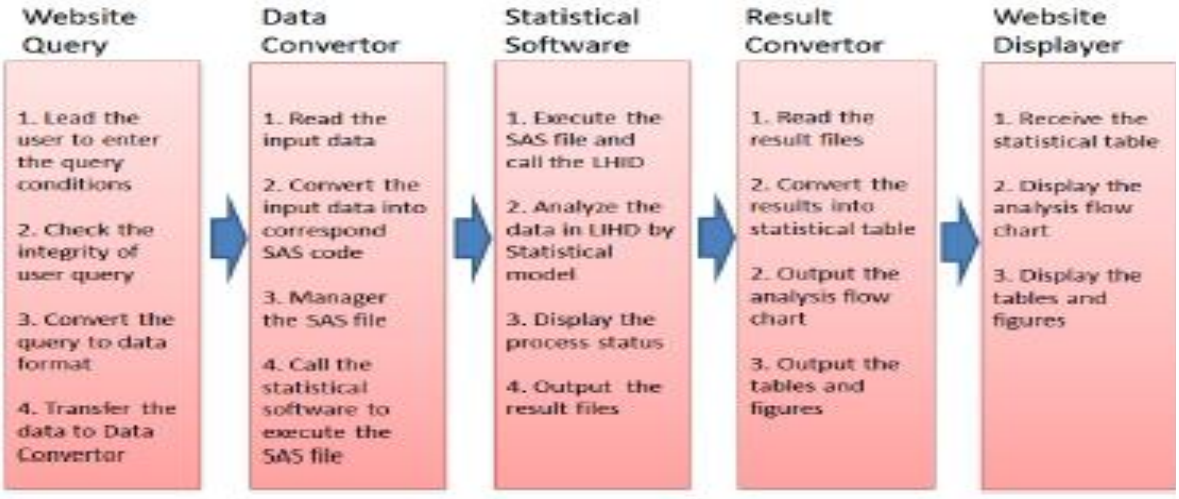


Figure 2. System flow of automatic analysis system

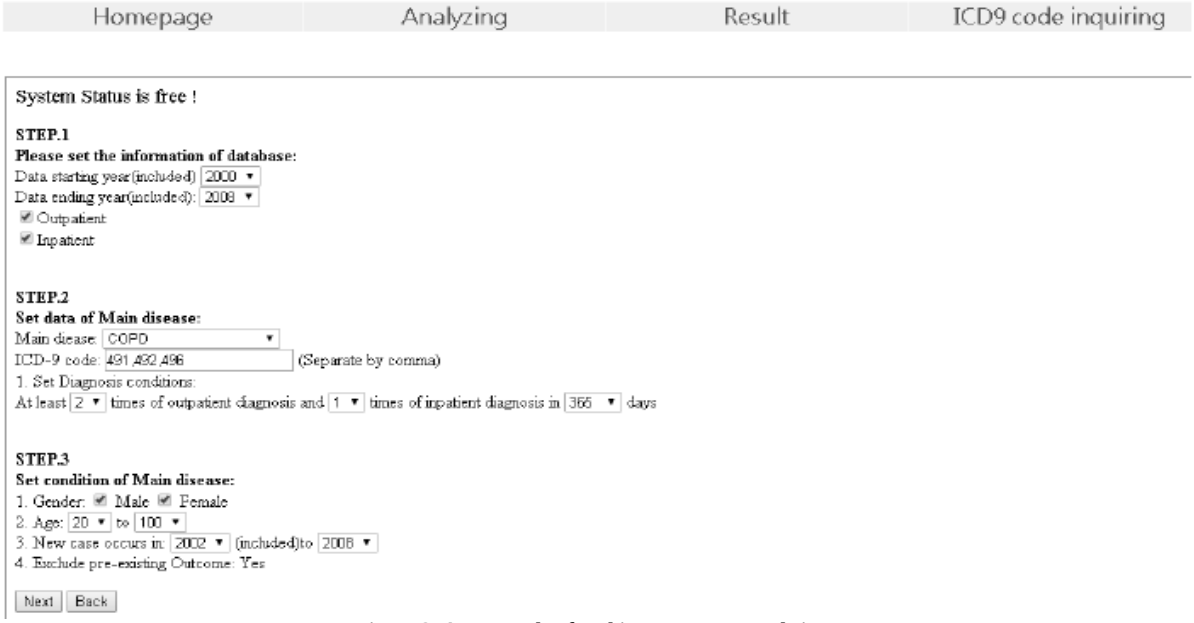


Figure 3. An example of making a query on website.

3.3Data Converter

This part is the core of the system. After receiving the query from the website, we transform the query into corresponding SAS code. We have chosen a specific statistical package, namely SAS statistical software, version 9.4 (SAS Institute Inc., Cary, NC), as our statistical software. SAS is able to manage the data from our data source as well as providing adequate statistical analysis functions.In Figure 4 we present an example of the transformation from a query into SAS code. As seen in Figure 4, we obtain original SAS code and transform this code into a module, which is able to generate SAS code via various input parameters. The module can be used repeatedly.

In Figure 5, the light green boxes indicate data and the orange boxes indicate data processing. All of the data derived from LHI database, both at the beginning and within a module is used to categorize the data into different groups; these include main disease, free of main disease, confounding factors and outcome. The first module also contains some sub-modules including a module for icd-9 diagnosis. We want our modules to be capable of being reused, so all modules should be suitable for variety of different groups.

Definitive diagnosis from the medical records in Figure 5 is important to avoid misdiagnosis. We have implement two ways to ensure a definitive diagnosis. The first one is we provide an option that allows the researcher to choose the number of outpatient/inpatient visits made by the subject because a diagnosis is accepted as definitive, while the other involves a time selection. The method we use in Figure 5 is to match based on a greedy algorithm; sex, age group and index date are considered when matching. Take 1:2 matching as instance; in this case each patient in study group is matched to two randomly-selected patients in the control group with the same status regarding sex, age group and the year of index date. If the control group

does not have enough patients, the study group will exclude some patients in order to maximize exact matches. In order to analyze the correlation between main disease, outcome and confounding factors, the incidence rate is calculated as the number of events per 1000 person-years. The baseline characteristics of the study group and control group are then compared by chi-square test and a p-value generated from chi-square test. Univariate and multivariate Cox regression hazard models are used to compute the crude hazard ratio (HR), the adjusted hazard ratio (aHR) and the 95% confidence interval (CI). We adjustment for a number of covariates, namely: main disease, sex and confounding factors. Statistical analysis is significant if the p-value is less than 0.05.

```
/*===== Calculate Age and delete unnecessary=====*/
DATA dm.maindisease_distinct_2_age;
SET dm.maindisease_distinct_1_2_sex;
 AA=input(FUNC_DATE,YYMMDD10.);
 FF=input(ID_BIRTHDAY,YYMMDD10.);
 format  AA YYMMDD10. FF YYMMDD10.;
 age=int(sum(AA-FF)/365);
if age < 20 or age >99 then delete;
run;
```

Transform a piece of SAS code into a module

```
public LinkedList<String> Patient_exclude_age(String status ,int start_age ,int end_age){

    LinkedList<String> list = new LinkedList<String>();
    list.add("");
    list.add("/*===== Calculate Age and delete unnecessary =====*/");
    list.add("DATA dm." + status + "_distinct_2_age;");
    list.add("SET dm." + status + "_distinct_1_2_sex;");
    list.add("  AA=input(FUNC_DATE,YYMMDD10.);");
    list.add("  FF=input(ID_BIRTHDAY,YYMMDD10.);");
    list.add("  format  AA YYMMDD10. FF YYMMDD10.;");
    list.add("  age=int(sum(AA-FF)/365);");
    list.add("if age < "+ start_age +" or age >" + end_age +" then delete;");
    list.add("run;");

    return list;

}
```

```
/*===== Calculate Age and delete unnecessary=====*/
DATA dm.outcome_distinct_2_age;
SET dm.outcome_distinct_1_2_sex;
 AA=input(FUNC_DATE,YYMMDD10.);
 FF=input(ID_BIRTHDAY,YYMMDD10.);
 format  AA YYMMDD10. FF YYMMDD10.;
 age=int(sum(AA-FF)/365);
if age < 50 or age >99 then delete;
run;
```

Set input parameters into module to generate SAS code
(status = outcome , start age =50 , end age = 99)

Figure 4 An example of transformation between query and SAS code.

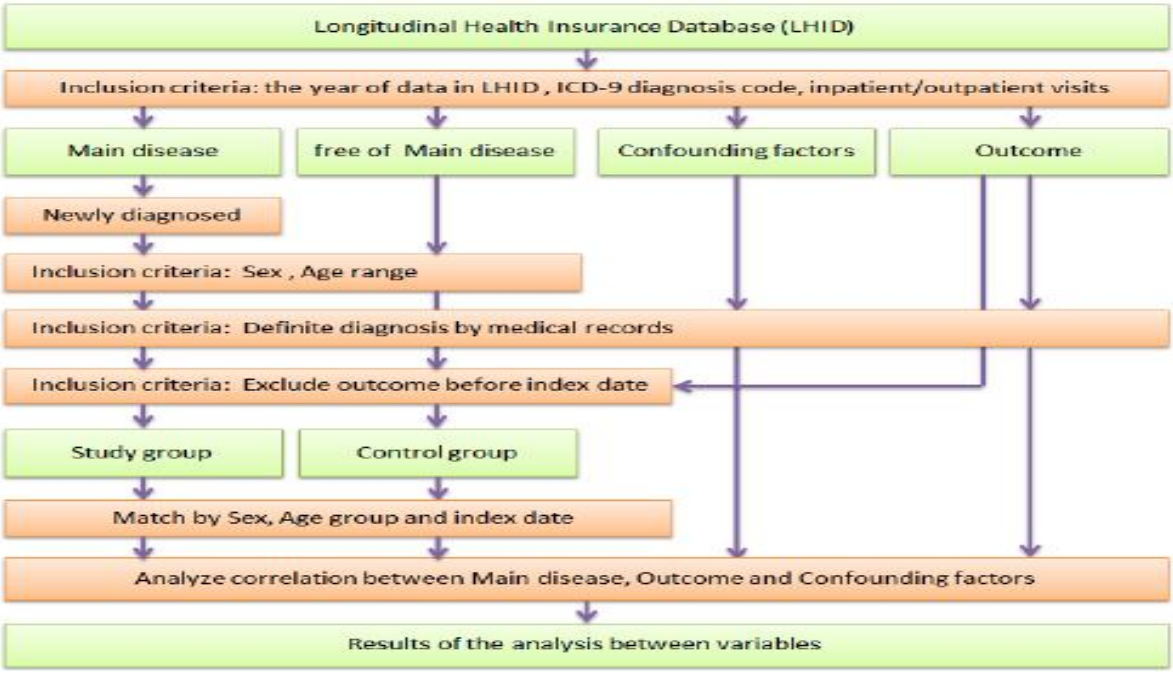


Figure 5. Flow of data processing for a cohort study (Light green boxes are data and orange boxes are data processing).

3.4Result Convertor and Website Displayer

In this part, we take the SAS output and convert the output files into result tables.The tables containing information on baseline characteristics in patients together with the risk ratios computed by Cox model, univariate analysis and multivariable analysis.

IV. RESULTS

4.1 Website of Our System

The website of our system consists of four parts: the homepage, a page for making a query to analyze the correlation between diseases, a page to display the results of analysis created by the system and a hyperlink to an online ICD-9 search engine. The users follow directions that help them make a query step by step. It is easily to input the parameters and this reduces the need for communication between researchers and the IT staff. We provide some default options to make entering the query faster. If a researcher selected a disease, we will automatically display the icd-9 code of the disease below it. Researcher can modify the diagnosis code to the code that they desired. We support three or five digits diagnosis codes such as 585 and 585.1. Our analysis system takes about one to two hours to carry out an analysis depending on the input parameters. Thus this is a fast and time saving approach to constructing a cohort study, carrying out coding and obtaining an analysis.

4.2 Verification of Our System

After building our system, we need to verify the results of the analyses created by our system. During verification, we used cases with three different origins: firstly, a situation where the relationship between two diseases is simply common sense, secondly, a situation where there are results that have been published in a journal and have used the same database and, thirdly, a situation were other researchers using a cohort study design has obtained a result but not using the Taiwan database. For fairness, we followed the input parameters of the previous researchers as much as it was possible to do.After building our system, we need to verify the results of the analyses created by our system. During verification, we used cases with three different origins: firstly, a situation where the relationship between two diseases is simply common sense, secondly, a situation where there are results that have been published in a journal and have used the same database and, thirdly, a situation were other researchers using a cohort study design has obtained a result but not using the Taiwan database. For fairness, we followed the input parameters of the previous researchers as much as it was possible to do.

4.3Verification by Common Sense

For this verification, we used three situations involving common sense to test our system. The first case was COPD and lung cancer. COPD and lung cancer are both smoking-related chronic diseases and 50% to 70% of patients with lung cancer have previously had an existing COPD diagnosis [16] and patients with reduced lung function are known to show an association with a higher proportion of lung cancer diagnoses [17]. Mannino et al. [17] showed that moderate/severe COPD patients have a high risk of developing lung cancer (HR = 2.8; 95% CI: 1.8-4.4) compared with patients with normal lung function. Table 1 shows the disease names for three case studies and includes the corresponding input parameters that we set. In first case, we chose data from the NHI database between 1997 and 2008 and the study group was newly diagnosed patients from 2001 to 2008. To reduce any potential misclassification of patients, a definitive diagnosis was needed for the corresponding patients and the input parameters are shown in Table 1. In this first case study, the result of the analysis by our system shows that COPD patients have a higher risk of lung cancer (crude HR = 4.099; adjusted-HR = 3.957) compared to non-COPD patients. Thus we have obtained similar results to previous research despite there being many differences between the two research approaches. The second case study is osteoporosis and bone fracture. Osteoporosis is a systemic skeletal disease that decreases bone strength and increases the risk of fracture [18]. The sequelae of osteoporosis include fractures and their complication and more than 2.3 million fractures are thought to be associated with osteoporosis every year in Europe and the USA [18],[19]. The results of the analysis is that patients with osteoporosis have an about 3.6-fold increased risk of suffering from a fracture (crude HR = 3.741; adjusted-HR = 3.637). The third case study is diabetes and gangrene of the foot. Diabetic foot is a complication of diabetes mellitus and is known to affect many patients with diabetes [20]. Wet gangrene is the most common type of type 1 diabetic foot complication [21]. We analyze the correlation between the two diseases and found that patients with diabetes had an about 5.7-fold increase in risk of suffering from foot gangrene (crude HR = 5.746; adjusted-HR = 5.764) compared with non-diabetic patients.

4.4 Verification by Journal

Next we choose a number of publications that used a similar research design to create cohort studies and also used the same database as our system, namely the NHIR database or the LHI database. We randomly selected ten articles from different journals, and compare the journals' results and the analytical results that our system outputted. We have outlined some details of the ten cases in Table 2; these are main disease, outcome, risk ratio, data selection, age and reference. The risk ratios include three common ratios and the confidence intervals (CI); the former are incidence rate ratio (IRR), crude hazard ratio (HR)

and adjusted hazard ratio (aHR). Data selections was based on the year of the data and year of onset, which means the time period over which the main disease occurred. Information on age was used to make sure that there were similar conditions for the study group between the original study and the analysis carried out by our system. Some parameters were directed introduced into the analysis system and these parameters were obtained from the published papers [6]-[15]. Due to lack of certain types of data, some variables, such as drugs usage, physical examination and surgery, cannot be introduced into our system.

Number	Main disease	Outcome	crude HR	adjusted HR	year of data	year of onset	age	definite diagnosis of main disease	definite diagnosis of outcome
1	COPD	Lung Cancer	4.099 (3.652-4.600)	3.957 (3.514-4.455)[a]	1997-2008	2001-2008	20-99	2 / 1 in 365 days[d]	2 / 1 in any time
2	Osteoporosis	Bone Fracture	3.741 (3.050-4.588)	3.637 (2.959-4.469)[b]	1997-2008	2001-2008	20-99	2 / 1 in any time	1 / 1 in any time
3	Diabetes	Foot Gangrene	5.746 (4.576-7.215)	5.764 (4.564-7.279)[c]	1997-2008	2001-2008	20-99	2 / 1 in 365 days	1 / 1 in any time

[a] COPD patients adjusted for some confounding factors: hypertension, diabetes, hyperlipidemia, heart disease, stroke, depression, arthritis, asthma, osteoporosis, liver disease, collagen vascular disease and anemia.

[b] Osteoporosis patients adjusted for some confounding factors: hypertension, diabetes, hyperlipidemia, obesity, liver cirrhosis, hemolytic anemia and spinal cord injury [10].

[c] Diabetes patients adjusted for some confounding factors: hypertension, hyperlipidemia, obesity, heart disease, cardiovascular disease (CVD), chronic kidney disease (CKD), coronary artery disease (CAD), depression, sleep disorder and cancer.

[d] Outpatient / inpatient visit times in a time period after index date of main disease.

Table 1 Three Cases Studies Based on Common Sense used to Verify the Results of Analysis System

Number	Main disease	Outcome	IRR	crude HR	adjusted HR[a]	year of data	year of onset	age	Reference
1	Rheumatoid Arthritis	Depression	-	1.72 (1.51-1.96)	1.74 (1.48-1.95)	1996-2011	2000-2004	≥18 years old	[6]
2	Schizophrenia	Peptic Ulcer	1.27 (1.06-1.52)	1.27 (1.07-1.50)	1.17 (0.97-1.41)[b]	1996-2009	1998-2001	mean age 36.5	[7]
3	COPD	Dementia	-	1.45 (1.39-1.51)	1.27 (1.20-1.36)	1996-2010	1998-2008	≥20 years old mean age 68.2	[8]
4	Osteoporosis	Dementia	1.71 (1.65-1.78)	-	1.46 (1.37-1.56)	-2010	2000-2010	≥50 years old mean age 67.5	[9]
5	Osteoporosis	Gallstone	-	1.50 (1.21-1.86)	1.35 (1.07-1.69)	1997-2010	2003-2005	≥50 years old	[10]
6	Sleep Disorders	Erectile Dysfunction	2.24 (2.00-2.51)	-	2.11 (1.89-2.37)	-2010	2002-2008	mean age 48.9	[11]
7	COPD	Peripheral Arterial Disease	2.34 (2.31-2.37)	-	1.48 (1.43-1.54)	-2010	1998-2008	≥20 years old mean age 69.9	[12]
8	Schizophrenia	Chronic Kidney Disease	1.36 (1.13-1.63)	-	1.25 (1.04-1.50)	-2010	2003-2007	≥18 years old	[13]
9	Migraine	SSNHL	1.80 (1.22-2.61)	-	-	2000-2009	2000-2009	20-64 years old mean age 38.5	[14]
10	Pulmonary Tuberculosis	Dementia	-	1.16 (1.02-1.32)	1.21 (1.06-1.38)	1996-2011	2000-2010	mean age 59.7	[15]

IRR, incidence rate ratio; HR, hazard ratio; COPD, chronic obstructive pulmonary disease; SSNHL, sudden sensor neural hearing loss
[a] Adjusted some confounding factors mentioned in reference.
[b] Statistical analysis is not significant.

Table 2 The Results from Some Journals used to Verify the Results of the Present Analysis System.

Number	Main disease	Outcome	definite diagnosis of main disease	crude HR[c]	adjusted HR[a][c]	IRR[b]	crude HR[a]	adjusted HR[a]	year of data	year of onset	age	Reference
11	Rheumatoid Arthritis	Depression	2 / 1 in any time	2.093 (1.830-2.393)	2.042 (1.785-2.336)	-	1.72 (1.51-1.96)	1.74 (1.48-1.95)	1997-2008	2000-2004	20-99 years old mean age 53.9	[6]
12	Schizophrenia	Peptic Ulcer	2 / 1 in 180 days	1.288 (1.138-1.460)	1.213 (1.072-1.373)	1.27 (1.06-1.52)	1.27 (1.07-1.50)	1.17 (0.97-1.41)[d]	1997-2008	1999-2001	0-99 years old mean age 38.5	[7]
13	COPD	Dementia	2 / 1 in 365 days	1.980 (1.875-2.090)	1.672 (1.583-1.766)	-	1.45 (1.30-1.51)	1.27 (1.20-1.36)	1997-2008	1999-2008	20-99 years old mean age 55.9	[8]
14	Osteoporosis	Dementia	2 / 1 in any time	1.856 (1.733-1.988)	1.784 (1.665-1.911)	1.71 (1.65-1.78)	-	1.46 (1.37-1.56)	1997-2008	2000-2008	50-99 years old mean age 67.2	[9]
15	Osteoporosis	Gallstone	2 / 1 in any time	1.433 (1.251-1.642)	1.359 (1.185-1.558)	-	1.50 (1.21-1.86)	1.35 (1.07-1.69)	1997-2008	2003-2005	50-99 years old mean age 67.0	[10]
16	Sleep Disorders	Erectile Dysfunction	2 / 1 in any time	3.204 (2.826-3.633)	2.938 (2.585-3.339)	2.24 (2.00-2.51)	-	2.11 (1.89-2.37)	1997-2008	2002-2008	20-99 years old mean age 46.9	[11]
17	COPD	Peripheral Arterial Disease	2 / 1 in 365 days	1.870 (1.669-2.097)	1.606 (1.432-1.802)	2.34 (2.31-2.37)	-	1.48 (1.43-1.54)	1997-2008	2002-2008	20-99 years old mean age 60.7	[12]
18	Schizophrenia	Chronic Kidney Disease	2 / 1 in any time	1.820 (1.362-2.431)	1.601 (1.198-2.149)	1.36 (1.13-1.63)	-	1.25 (1.04-1.50)	2000-2008	2003-2007	20-99 years old mean age 40.6	[13]
19	Migraine	SSNHL	2 / 1 in any time	1.742 (1.151-2.636)	1.664 (1.096-2.527)	1.80 (1.22-2.61)	-	-	2000-2008	2002-2008	20-69 years old mean age 42.5	[14]
20	Pulmonary Tuberculosis	Dementia	2 / 1 in any time	1.294 (1.150-1.455)	1.213 (1.078-1.366)	-	1.16 (1.02-1.32)	1.21 (1.06-1.38)	1997-2008	2000-2008	20-99 years old mean age 60.1	[15]

IRR, incidence rate ratio; HR, hazard ratio; COPD, chronic obstructive pulmonary disease; SSNHL, sudden sensor neural hearing loss.
[a] Adjusted some confounding factors mentioned in references [6-15].
[b] The risk ratios come from corresponding references [6-15].
[c] The risk ratios are generated by analysis system and statistical analysis is all significant in risk ratios.
[d] Statistical analysis is not significant.
p-value < 0.05.

Table 3 The Results of the Analysis System using Similar Parameters Input in order to Verifying the System

Table 3 is a result table generated by our analysis system with similar input parameters to those of the corresponding reference [6]-[15]. Table 2 corresponds to Table 3, number 1 in Table 2 corresponds to number 11 in Table 3, number 2 in Table 2 is corresponds to number 12 in Table 3 and so on.

4.5The results of the correlation between main disease and outcome

In Table 2, the risk ratios for numbers 1 to 10 are all significant except the adjusted HR of number 2 (aHR=1.17; CI: 0.97-1.41). The risk ratios in Table 3 are all significant, and all correlations show that the main disease increased the risk of developing the targeted outcome. In numbers 2 and 12, the crude HRs are similar and the two confidence intervals of adjusted HR overlap (CI: 0.97-1.41 and 1.072-1.373), but the adjusted HR of number 2 is not significant. One of the main reasons that may have produced some differences is that more variables were added and adjusted. The research in number 2 has more variables and these include urbanization and monthly income. This is good way of reducing differences between groups but our system does not implement a function that allows analysis of urbanization and income. The second reason is that the onset patients were collected from 1998 to 2001, while our onset patients were collected from 1999 to 2001. Our dataset lacks data for 1996, 2009 and 2010, we decided to study patients from 1999 to 2001. Taking the above into account, we believe that the differences in the results for Number 2 and 12 are quite reasonable.

4.6 Some results comparing the impact of definitive diagnosis

Definitive diagnosis of the main disease has a major impact on the results of an analysis. In numbers 1 and 11, the results indicate that rheumatoid arthritis (RA) can lead to depression, and both adjusted HR are significant. The major difference between Numbers 1 and 11 is the definitive diagnosis of the main disease. Lin et al. [6] (number 1) chose RA patients with a primary diagnosis of RA; however, our system does not provide a primary diagnosis only. The impact of using a primary diagnosis is to make the number of patients with rheumatoid arthritis fewer than in our study group, but the gender ratio is similar at about 65%. We also think that the ratios of comorbidities are different because of the definitions used for the comorbidities. Numbers 5 and 15 contain the same definite diagnosis difference, this time for osteoporosis patients; this is because receiving at least one bone mineral density examination was central to the research of Klahan et al. [10]. In our verification, various other analytical result tables have some similarities and differences when comparedto the results of previous research [9],[11],[13],[15],[16]; nonetheless, the similarities are greater than the differences. We have not

compared risk factors using univariable and multivariable analysis because it is very difficult to obtain the same definite diagnosis for all comorbidities. There is a need to collect all the definite diagnoses of comorbidities or, alternatively, to ask the authors of the earlier study for the parameters used.In our verification, various other analytical result tables have some similarities and differences when compared to the results of previous research [9],[11],[13], [15],[16]; nonetheless, the similarities are greater than the differences.

DISCUSSION

In our study, we have described the implementation of an automated analysis system that increases efficiency when carrying out cohort study design research based on the LHID. The results show that our analysis system automatically finishes the queries gathered from researchers and actually carries out cohort study research faster than the traditional approach that involves coding by IT staff. The average execution time of a query is about 1-2 hours and this is highly acceptable. We have not focused on obtaining better performance by the system because this is another topic and would involve optimization of the cohort study designs created by the system.

We have shown here how we designed the analysis system and presented details of the system design. A similar study design, for example targeting case-control studies, should be easy to implement in the same way. However, in such a system it is likely that would be too many variables that not are available, which would affect the reliability of the system, and therefore we have only focused on the correlations between pairs of diseases. Nonetheless, some variables are still difficult to implement when analyzing pairs of diseases; these include drug usage and medicament dosage. In order to obtain a balance between design problems and the functionality of system, we have retained various common variables and important parts of the data procedure as shown in Figure 5. We think that it should be possible to increase the functionality of our system with the next version.

We have verified the correlation results obtained using our automated analysis system as well as some of the specific results generated by our system. Each correlation analyzed by the system was found to be quite similar to previous studies and all were significant [6]-[15],[17]. The system seems to be reliable in terms of the correlation results generated by our system, and other results are also quite similar. We think that the results generated by our system will help to accelerate research when two disease (a main disease and an outcome) are being investigated by researchers and it should be possible to conduct quickly a more detailed analysis after the system has initially created an analysis of the research question.

There are several limitations that affect the database used and the system itself. First, we do not have information on lifestyle, on family history, on various areas of clinical data and on the severity of a given disease. These are confounding factors and may impact on the results of the analysis. Second, in order to reduce effect of inaccurate diagnosis codes, definitive diagnosis of disease is needed, for example, using at least two diagnoses of the same disease; this could be problematic with some diseases. Third, all the disease diagnoses are based on the patients' medical records and thus we may underestimate the incidence of a given disease in the general population [11]. Finally, verification of the results obtained by our system is very important and the new medical results generated by our system need to be further confirmed by professional medical researchers independently by comparing the systems results with other similar research. The major strength of our study is that it is based on the LHID, which is very large database of Taiwan’s complete population, and the statistical results are thus supported by the very large size of this database. In conclusion, our study is first to implement an automated analysis system to analyze the correlation between diseases using a cohort study design and the LHID. The system is very efficient and it should make research that uses correlations between diseases much faster and more efficient. Researchers will be able to conduct more detailed analysis more quickly in the future. In terms of results, our analysis system is able to provide similar results to those obtained by other cohort studies, and the correlation between diseases obtained are generally in agreement with those of earlier studies. In the future, we will add more variables, examples being operation code, urbanization and monthly income, to our system, which should help to produce more accurate analysis results and then we will seek help from medical researchers to confirm further any novel results.

REFERENCES

[1] National Health Insurance Research Database [Online]. Available: http://nhird.nhri.org.tw/

[2] Y.-C. Chen, H.-Y. Yeh, J.-C. Wu, I. Haschler, T.-J. Chen, and T. Wetter, "Taiwan's national health insurance research database: Administrative health care database as study object in bibliometrics," Scientometrics, vol. 86, pp. 365-380, 2011.

[3] N. R. Coordinators., "Database resources of the National Center for Biotechnology Information," Nucleic acids research 42.Database issue, D7, 2014.

[4] K. B., "Epidemiology and Biostatistics: An Introduction to Clinical Research," Springer Science & Business Media, 2009.

[5] J. W. Song, Chung, K.C., "Observational studies: Cohort and case-control studies," PlastReconstrSurg, vol. 126, pp. 2234–2242, 2010.

[6] G. H. Lin MC, Lu MC, Livneh H, Lai NS, Tsai TY., "Increased risk of depression in patients with rheumatoid arthritis: a seven-year population-based cohort study," Clinics, vol. 70, pp. 91-96, 2015.

[7] C. C. Liao CH, Chang SN, et al., "The association of peptic ulcer and schizophrenia: a population-based study," J Psychosom Res, vol. 77, pp. 541-546, 2014.
[8] L. C.-L. Liao W-C, Chang S-N, Tu CY, Kao CH., "The association between chronic obstructive pulmonary disease and dementia: a population-based retrospective cohort study," Eur. J. Neurol., vol. 22, pp. 334-340, 2015.
[9] C. C. Chang KH, Lin CL, Sung FC, Wu TN, Kao CH., "Increased risk of dementia in patients with osteoporosis: a population-based retrospective cohort analysis," Age, vol. 36, pp. 967-975, 2014.
[10] S. Klahan, et al., "Osteoporosis increases subsequent risk of gallstone: a nationwide population-based cohort study in Taiwan," BMC gastroenterology, vol. 14, p. 192, 2014.
[11] H. H. Lin, et al., "Increased risk of erectile dysfunction among patients with sleep disorders: a nationwide population□based cohort study," International journal of clinical practice, 2015.
[12] T.-C. Shen, et al., "Chronic Obstructive Pulmonary Disease is Associated with an Increased Risk of Peripheral Arterial Disease," 內科學誌, vol. 25, pp. 272-280, 2014.
[13] H. Y. Tzeng NS, Ho SY, Kuo YC, Lee HC, Yin YJ, et al., "Is schizophrenia associated with an increased risk of chronic kidney disease? A nationwide matched-cohort study," BMJ open, vol. 5, p. e006777, 2015.
[14] L. C. Chu CH, Fuh JL, Shiao AS, Chen TJ, Wang SJ., "Migraine is a risk factor for sudden sensorineural hearing loss: a nationwide population-based study," Cephalalgia, vol. 33, pp. 80-86, 2013.
[15] Y.-H. Peng, et al., "Increased Risk of Dementia Among Patients With Pulmonary Tuberculosis A Retrospective Population-Based Cohort Study," American journal of Alzheimer's disease and other dementias, vol. 30, pp. 629-634, 2015.
[16] R. I. Yao H, "Current concepts on the role of inflammation in COPD and lung cancer," CurrOpinPharmacol, vol. 9, pp. 375-383, 2009.
[17] A. S. Mannino DM, Petty TL, Redd SC., "Low lung function and incident lung cancer in the United States: data from the First National Health and Nutrition Examination Survey follow-up," Arch Intern Med, vol. 163, pp. 1475-1480, 2003.
[18] W. H. Organization., "Prevention and Management of Osteoporosis," Technical Report Series, No. 921, 2003.
[19] d. B. S. Cosman F, LeBoff MS, et al., "Clinician's Guide to Prevention and Treatment of Osteoporosis,"OsteoporosInt, vol. 25, pp. 2359-2381, 2014.
[20] T. Tan, Shaw, EJ, Siddiqui, F, Kandaswamy, P, Barry, PW, Baker, M, Guideline Development Group, "Inpatient management of diabetic foot problems: summary of NICE guidance," BMJ., vol. 342, pp. 702-703, 2011.
[21] V. K., "Evaluation of diabetic foot complications according to Amit Jain's classification," Journal of clinical and diagnostic research: JCDR, vol. 8, NC07, 2014.

★★★